From 2D to Stereoscopic-3D Visual Saliency: Revisiting Psychophysical Methods and Computational Modeling

Junle Wang

l'université nantes angers le mans
PÔLE DE RECHERCHE ET D'ENSEIGNEMENT SUPÉRIEUR

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Junle Wang

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université de Nantes
sous le label de l'Université de Nantes Angers Le Mans

Discipline : Automatique et Informatique Appliquée
Spécialité : Traitement du signal et des images
Laboratoire : Institut de Recherche en Communications et Cybernétique de Nantes (IRCCyN)

Soutenue le 16 Novembre 2012

École doctorale : STIM
Thèse n° : ED 503-181

From 2D to Stereoscopic-3D Visual Saliency: Revisiting Psychophysical Methods and Computational Modeling

Saillance Visuelle, de la 2D à la 3D Stéréoscopique : Examen des Méthodes Psychophysique et Modélisation Computationnelle

JURY

Rapporteurs :	**M^me^ Anne GUÉRIN DUGUÉ**, Professeur, Université Joseph Fourier **M^me^ Ingrid HEYNDERICKX**, Professeur, Delft University of Technology
Examinateurs :	**M^me^ Luce MORIN**, Professeur, Institut National des Sciences Appliquées **M. Fédéric DUFAUX**, Directeur de recherche CNRS, TELECOM ParisTech **M. Vincent RICORDEL**, Maître de conférences, Université de Nantes **M. Patrick LE CALLET**, Professeur, Université de Nantes
Directeur de thèse :	**M. Patrick LE CALLET**, Professeur, Université de Nantes
Co-encadrant de thèse :	**M. Vincent RICORDEL**, Maître de conférences, Université de Nantes

Acknowledgement

Looking back the time of pursuing a PhD, it has been such a great and unforgettable journey in my life. I am deeply grateful for all that I have received during this journey which has been full of challenge and pleasure. It is my great pleasure to take this opportunity to thank the people for their help and support over the past years.

First of all, I would like to express my most sincere gratitude to my dear supervisors Prof. Patrick Le Callet and Prof. Vincent Ricordel for offering me the opportunity to pursue my doctoral studies under their supervision. I appreciate all their time and support which make my PhD experience very enjoyable and productive. More importantly, I want to thank them for giving me the opportunity and freedom to pursue my personal research interest. I also want to thank them for helping me to adapt to the life in Nantes, and their continuous help and supports whenever I encounter any problem in France. They are nice, patient, intelligent, professional, and with charming personality. I have learned a lot from them. I would like to say, I am so lucky to have them as my supervisors and work with them in the past years.

I would also like to express my great appreciation to my colleagues Prof. Matthieu Perreira Da Silva, Prof. Marcus Barkowsky, as well as Prof. Damon M. Chandler from Oklahoma State University. I really thank them for sharing their professional expertise and making my PhD experience productive. A very special "thanks!" goes to Prof. Tusheng Lin from South China University of China, who encouraged me to start my career in scientific research and pursuit my PhD degree.

My gratitude goes as well to my collaborators and friends, in particular to Hantao Liu from Delft University of Technology, Ulrich Engelke from Philips Research and Sylvain Tourancheau from Mid Sweden University. The collaborations with them are productive and pleasurable. I also appreciate the excellent example they have provided as being a young researcher : intelligent and full of passion. Many thanks goes also to my colleges Romuald Pepion and Romain Cousseau for their continuous help and support in every possible way for the issues related to subjective experiments. It was them who made the difficult experiments efficient. Special thanks goes also to Prof. Zhou Wang from University of Waterloo for sharing his experience in scientific research with me.

Furthermore, I am very honored to have such highly renowned and competent experts in my Ph.D. committee. My great appreciation goes to Prof. Ingrid Heynderickx from Delft University of Technology, Prof. Anne Guérin Dugué from Université Joseph Fourier,

Prof. Luce Morin from INSA and Dr. Frédéric Dufaux from TELECOM ParisTECH.

My journey towards the Ph.D. would not have been possible without the love and encouragement of my family. I really appreciate the love and supports from my parents, my grandparents, and my uncle. I am also grateful for the continuous support from friends who shared the past years in Nantes with me, in particular from dear Pangzhi, Bangge, Kunshu, Tamara, Jiefu, Jiazi, Chenwei, Dingbo, He Hong Yang, Zhujie, Baobao, Chuanlin, Fengjie, Xiebo, Wenzi, Zhangyu, Yuwei, Biyi, Shuangjie, Shujin, Wangyu, and so on, I would like to thank you all so much.

Contents

Chapter 1

Introduction

1.1 Context

In everyday life, we are constantly receiving an abundant amount of information through various sense. Among the senses, sight is considered to be the most dominant one as it is comparably stronger developed than some of the other senses, such as smell and taste [Wandell 95]. However, our sensory system for vision, the human visual system (HVS), continually receives a really large amount of visual data ($10^8 - 10^9$ bits per second) [Borji 12]. This considerably larger amount of data is beyond our brain's capability to process all of them. To cope with this large amount of information, visual attention is one of the most important mechanisms deployed in the HVS to reduce the complexity of the analysis of visual scene [Wolfe 00]. Driven by visual attention, viewers can selectively focus their attention on specific areas of interest in the scene.

In the last decades of years, extensive efforts have been dedicated in the study of visual attention. Neurologists, psychologists, vision scientists, and computer scientists have taken part in and contribute to various aspects of visual attention. These efforts from different disciplines make the research of visual attention become a highly interdisciplinary field; different relevant disciplines deal with the research on visual attention from different points of view, and profit from each other.

1.2 Motivations

In recent years, the deployment of visual attention mechanism in image processing system has found increasing interest by computer scientists. Taking into account visual attention information becomes an effective way for improving various algorithms existing in image processing. Variety of areas, including compression [Park 02], retargeting [Wang 11a], retrieval [Vu 03], quality assessment [Liu 11], have been beneficial by being provided the information of the location in visual scene that attracts viewer's attention.

To fully exploit the benefits of visual-attention-based algorithm, the regions of a visual scene that attract attention need to be computationally identified, for which purpose computational visual attention models are developed and deployed. This is the reason that an increasing amount of efforts are being dedicated in the study of visual attention, particularly the studies of computational modeling of visual attention. As we know, the development of computational visual attention models largely rely on a particular issue: the ground truth. Therefore, the reliability of the ground truth is of particular importance.

So far, many studies rely on the fixation density maps (FDM). There are two reasons: (1) it is believed that there is a strong link between overt visual attention and eye movements; and (2) these FDM are obtained from eye-tracking experiment so they are believed to be reliable ground truth. The authenticity of the first reason has been shown in many studies in the literature, however, the certainty of the second reason remains an open question. As we know, eye-tracking experiment are usually time consuming and expensive. Therefore, several eye-tracking FDM databases have been made publicly available. However, these experiments were conducted independently in different laboratories even different countries. The outcomes of the experiments hence might depend on different factors related to the observer and the experimental design. It remains as an open question to the society how much we can trust this ground truth, and what is the impact of the difference between various experiments on image processing applications.

When visual attention is taken into account by the signal-processing community, the two terms, "salience" and "importance", have traditionally been considered synonymous. It is true that both of visual salience and visual importance denote the most visually "relevant" parts of the scene. However, from the vision scientist's point of view, they are two different concepts, since they come from two different mechanisms of visual attention: bottom-up and top-down. The two mechanisms are driven by different types of stimuli, and are formed in different visual pathways that go through different areas of the brain. Therefore, it would be worth to identify the two terms in the context of image processing. An even better way would be to quantify the relationship between them.

In recent years, another problem that researchers in the field of visual attention have to face is the impact of 3D. During viewing 3D content, depth perception of the scene is enhanced. This change of depth perception also largely changes human viewing behavior [Hakkinen 10, Huynh-Thu 11a]. Because of the emergence of 3D content and recent availability of 3D-capable display equipments, studies related to 3D visual attention have been gaining an increasing amount of attention in the last few years. Nevertheless, the community of 3D visual attention faces some problems:

(1) There are a relatively small number of models. Compared to the amount of computational models for 2D image/video, only a small number of models of 3D visual attention can currently be found in the literature. Developing effective 3D models, which can be for general purpose or specific condition, is of particular importance. However, developing 3D visual attention still faces some problems, e.g. the selection

and extraction of features, the pooling strategy for different feature channels, and the increased computational complexity due to the additional depth information.

(2) The ground truth is still lacking. As introduced above, ground truth plays a crucial role in the development of computational model. However, any published eye-tracking database of 3D images is still lacking in the community. The lack of ground truth leads to the difficulties of quantitatively assessing and comparing the performance for the existing 3D computational models.

(3) The impact of the enhanced depth perception on visual attention still needs to be further investigated. This knowledge can help in the computational modeling of 3D visual attention. However, several studies in the literature draw contradictory conclusions.

(4) Center-bias has been investigated for 2D viewing condition, it has also been demonstrated to have a large added value in predicting 2D saliency map. However, it remains an open question how center-bias varies with viewing condition and how it should be integrated in modeling 3D visual attention.

(5) New applications of visual attention (in 3D) need to be proposed and verified. So far, many areas in image processing have benefited from being provided with information of saliency. Nowadays, researchers have proposed that it might be possible to take advantage of visual attention to solve some intrinsic problem of 3D display techniques, e.g., the visual discomfort and visual fatigue caused by the conflicts between depth cues. However, the feasibility and the method to implement it still remain as open questions.

In order to provide some efforts for solving the problems mentioned above, we have done some works, which will be presented in the remainder of this thesis. The main content and the structure of the thesis is shown in the next section.

1.3 Outline of the thesis

In Chapter 2, we provide the reader with important concepts regarding visual attention, eye movements and the technique of eye-tracking. We also present some widely used models of visual attention for 2D content. Following Chapter 2, the remainder of the thesis is then structured into two parts.

The first part of the thesis includes three works that are concerned with the ground truth, i.e. the fixation density map (FDM), used in the study of computational modeling of visual attention. They can be considered to be regarding: (1) the accuracy of FDM, (2) the meaning of FDM, and (3) the creation of new FDM database.

- In Chapter 3, we investigate the reliability of eye-tracking data from different databases. Being a part of an international cooperation, we evaluate the similarity and difference between the FDM databases obtained from three eye-tracking experiments which are conducted in three different laboratories (and countries). We further study the impact of the inter-laboratories difference on being ground truth for image processing applications.

- In Chapter 4, we present a psychophysical study for quantifying the relationship between eye-tracking data and the result of a scoring experiment which is conducted for identifying regions-of-interest. In our study, the FDM are created by a free-viewing eye-tracking experiment; they are believed to provide information of bottom-up visual salience. On the other hand, the importance map (i.e. ROI) is believed to be linked with top-down mechanisms, which are more related to the meaning and the gist of the scene. Therefore, the study presented in this chapter can be considered as a study of the relationship between bottom-up and top-down mechanisms.

- In chapter 5, we present a binocular eye-tracking experiment. We also study several challenges (e.g. the acquisition of depth information, the window violation, and the adjustment of fixation's location according disparity) of conducting eye-tracking experiments for stereoscopic 3D content. Based on the eye-tracking experiment, we create a database containing eighteen stereoscopic 3D images of natural content, as well as their corresponding disparity maps and free-task viewing eye-tracking data. This database helps in solving the problem of lacking ground truth in the research area of 3D visual attention modeling.

The second part of the thesis includes five works focusing on computational modeling of visual attention, in particular on the modeling of 3D visual attention.

- In Chapter 6, we focus on the impact of perceived depth on the deployment of visual attention in 3D scene. We try to determine if there exists a so-called "depth-bias" in the viewing of 3D content on planar stereoscopic 3D display.

- In Chapter 7, we propose a new computational model of visual attention for stereoscopic 3D still image. We first introduce several state-of-the-art 3D visual attention models. Based on the way of using depth information, we propose a new taxonomy of existing computational models of 3D visual attention. Next, we propose a model which takes depth as an additional visual dimension. The measure of depth saliency is derived from the eye movements data which are obtained from an eye-tracking experiment using synthetic stimuli. Two different ways of integrating depth information in modeling 3D visual attention are then proposed and examined.

- In Chapter 8, we present a comprehensive study on the differences of center-bias in the 2D and 3D viewing conditions. We also propose and evaluate two different ways of integrating center-bias into 3D visual attention modeling. Finally, we propose a hybrid computational model of 3D visual attention taking into account both the depth saliency as well as center-bias.

- In Chapter 9, we propose a potential application of the 3D visual attention model. We quantify how two depth cues, defocus blur and binocular disparity, interactively

affect perceived depth. We propose that, by knowing the region of interest, it is possible to blur out parts of the scene in order to change the perceived depth of the scene, and further increase the quality of experience of 3D viewing.

At the end of this thesis, we present a conclusion of the whole thesis in Chapter 10. A summary of the contribution and some perspectives are also presented in this chapter.

Chapter 2

Visual attention: concepts, measurement, and modeling

It would be difficult to go directly into the specific studies without a general introduction of the background knowledge of visual attention. In this chapter, we firstly introduce concepts of visual attention as well as various mechanisms of attention (Section 2.1). Secondly, we present a brief introduction of the HVS in Section 2.2. Next, in Section 2.3, we present an introduction of different types of eye movements as well the technique of measuring eye movements, i.e. the eye-tracking. Finally, we introduce some typical state-of-the-art computational models of visual attention in Section 2.4. An conclusion of this chapter is presented at the end of the chapter in Section 2.5.

2.1 Concepts of visual attention

The oldest and most famous definition of attention, which is provided by the psychologist William James [James 80], dates back to year 1890:

"Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others.".

In HVS, attention plays an important role in visual processing by keeping the essential visual information. Tsotsos et al. [Tsotsos 95] proposed that visual attention is a mechanism having at least the following basic components: (1) the selection of a region of interest in the visual field; (2) the selection of feature dimensions and values of interest; (3) the control of information flow through the network of neurons that constitutes the visual system; and (4) the shifting from one selected region to the next in time. Driven by visual attention, viewers can therefore selectively focus their attention on specific areas of interest in the scene.

2.1.1 Overt attention and covert attention

There are two types of attention, namely overt attention and covert attention. These two types of attention are differentiated base on their relation with eye movements.

- Overt attention is usually associated with eye movements. This type of attention is easy to observe: when we focus our attention to an object, our eye moves to fixate this object. One of the earliest studies of overt attention came from Yarbus [Yarbus 67]. He studied the correlation between visual attention and the eye movements during the viewing of human face (see Figure 2.1.1).
- In addition to overt attention, William James [James 80] found that human are able to attend to peripheral locations of interest without moving the eyes; this type of attention is named as covert attention. An advantage of covert attention is its independence of motor commands [Frintrop 06]. Since the eyes do not need to be moved to focus attention on a certain region, covert attention is much faster as compared to overt attention. An example of covert attention is driving, where a driver keeps his eyes on the road while simultaneously covertly monitoring the status of signs and lights [Borji 12].

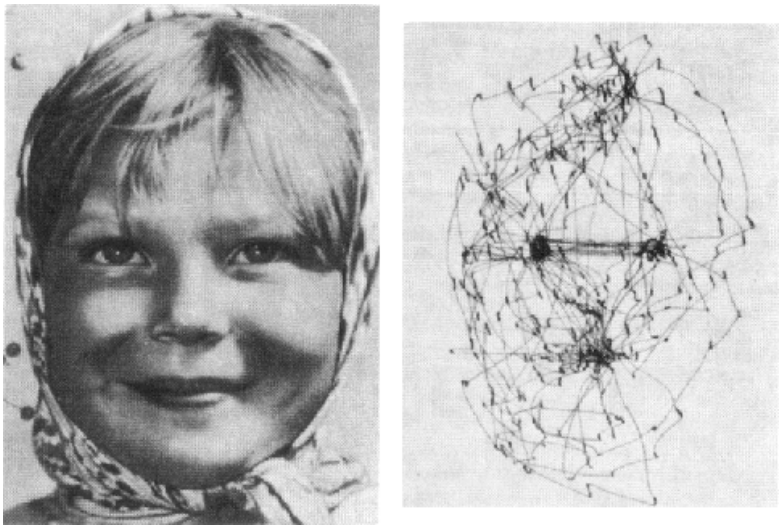


Figure 2.1.1: Gaze tracks during the observation of a face [Yarbus 67].

Overt attention and covert attention are not independent. Human cannot attend to one location while moving their eyes to a different location [Deubel 96]. The covert shift of attention to a location is linked to eye movement by setting up a saccade to that location [Peterson 04].

Most of current studies, specially the studies of computational modeling of visual attention, are with respect to overt attention, since overt attention can be measured in

a straightforward way by using eye-tracking. However, it is difficult to measure covert attention. The computational framework for covert attention is also still lacking.

2.1.2 Bottom-up attention and top-down attention

A shift of attention can be caused by two categories of cues: one is referred to as bottom-up cue, and the other one is referred to as top-down cue.

Bottom-up attention is driven by the characteristics of a visual scene, i.e. the bottom-up cues. Bottom-up attention is hence also referred to as stimulus-driven attention or exogenous attention. Bottom-up attention is fast, involuntary, and most likely feed-forward [Borji 12]. Since bottom-up attention is usually driven by low-level features (e.g. intensity, color, and orientation), in order to attract human's bottom-up attention, an area must be sufficiently distinctive compared to the surrounding area with respect to these low-level visual features.

On the other hand, top-down attention is based on "higher level" information, such as knowledge, expectations and current goals [Desimone 95]. Top-down attention is thus also referred to as concept-driven attention, goal-driven or endogenous attention. As compared to bottom-up attention, top-down attention is slow, voluntary and driven by the task demands. A famous illustration of top-down attention comes from Yarbus's work in 1967 [Yarbus 67]. He demonstrated how eye movements varied depending on the question asked to the subject during observing the same scene (see Figure 2.1.2).

2.1.3 Feature Integration Theory

One of the best known and most accepted theories of visual attention is the "Feature Integration Theory", which was proposed by Treisman [Treisman 80] at the beginning of the 1980s. This theory has been the basis of many computational models of visual attention.

In [Treisman 80], Treisman claimed that "different features are registered early, automatically and in parallel across the visual field, while objects are identified separately and only at a later stage, which requires focused attention". According to the Feature Integration Theory (FIT), the different features of stimuli are firstly encoded in areas partially independent. In addition, our hierarchical cortical structures are organized in order to make the detection of these features relatively independent of their positions in the visual scene.

The FIT introduced a concept of "feature maps", which are topographical maps that highlight salience according to the respective feature. Information of the feature maps ia then collected in a "master map of location". This map indicates the location of the objects, but does not provide information about what the objects are.

Finally, to construct a coherent representation of the scene, selective attention is used. The scene is scanned by an attentional beam of variable size (see Figure 2.1.3). This beam blocks the information which is not located within its radius. It is thus possible

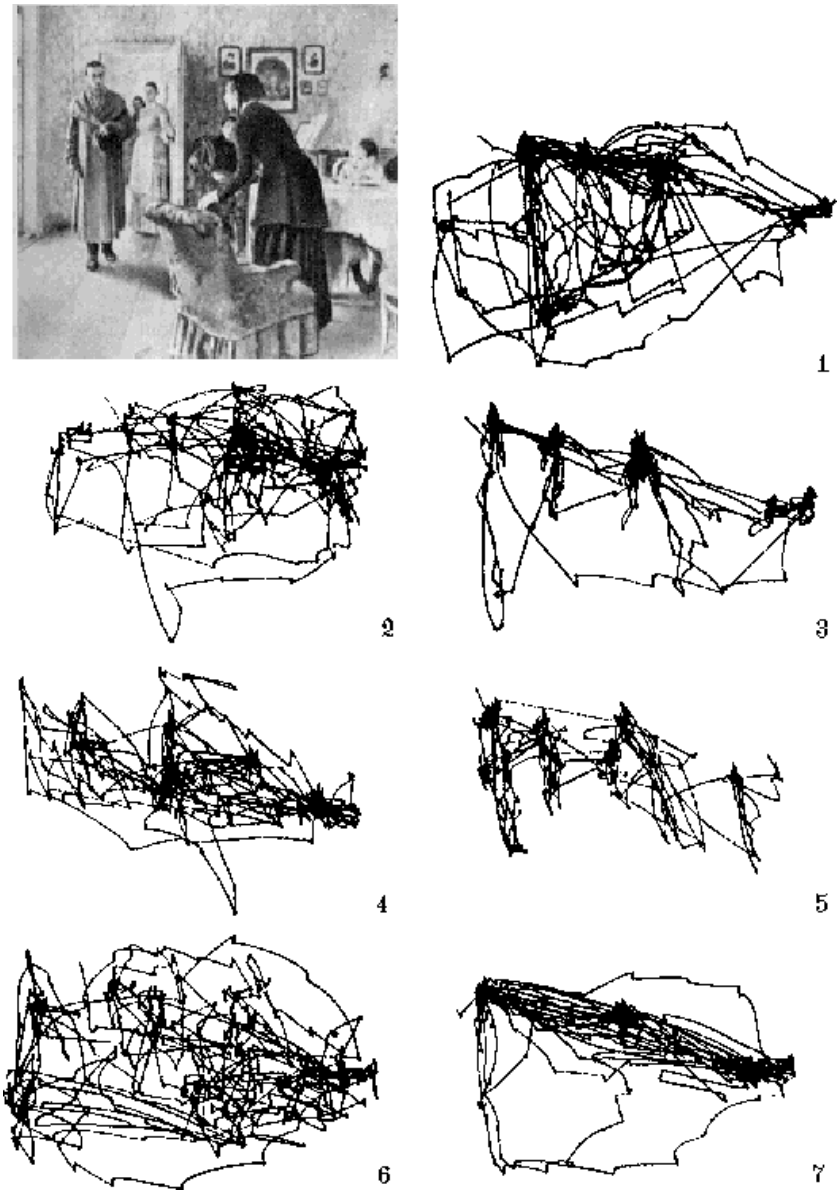


Figure 2.1.2: An example that eye movements depend on observer's viewing task [Yarbus 67]. (1) No question was asked. (2) Judge economic status of the family. (3) What is the age of each person. (4) What were they doing before the visitor arrived. (5) Memorize what clothes they are wearing. (6) Memorize the positions of the people and the objects in the scene. (7) Estimate how long the unexpected visitor had not seen the family. In this experiment, each recording of eye movements lasted for 3 minutes.

to match all the features found in this area in order to build a coherent representation. By moving the beam over time, our brain construct gradually a global perception of the scene.

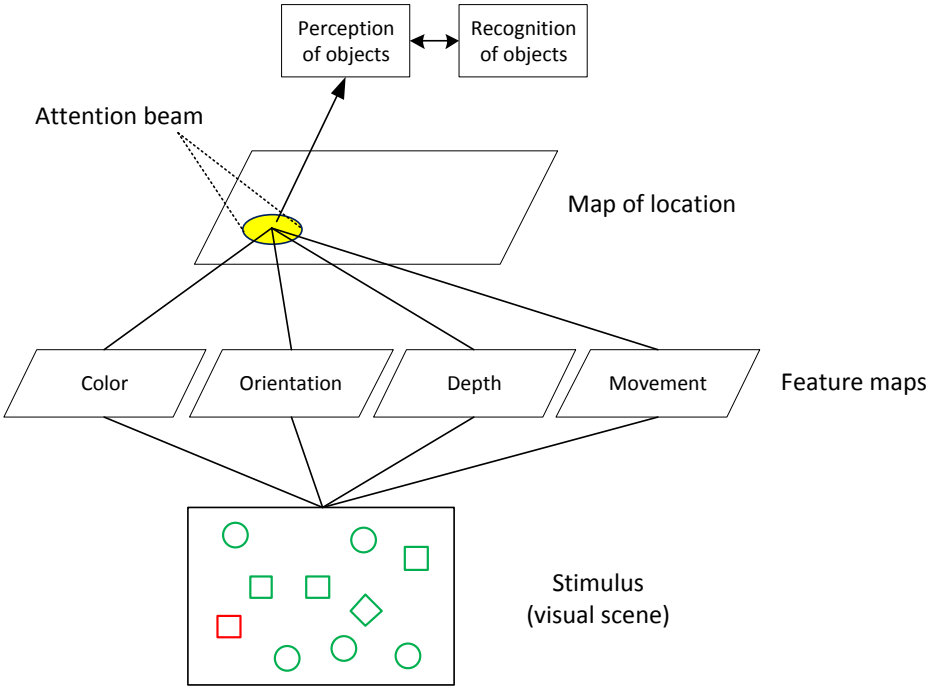


Figure 2.1.3: Illustration of Feature Integration Theory (FIT). [Perreira Da Silva 10]

2.2 Visual attention and human visual system

Since visual attention is a mechanism involved in the process of visual perception, it is of importance to introduce also the knowledge regarding how visual information is processed in the human visual system. While being far from an exhaustive explanation of the HVS and the mechanisms involved in the processing of visual information, we present briefly in this section an introduction of retina and different areas of the visual cortex (Figure 2.2.1) that allow determine the main characteristics of the HVS.

Retina

The retina is a light-sensitive surface, which has over 100 million photoreceptor cells [Mather 09]. The photoreceptor cells are responsible for transducing light energy into neural signals. Note that the retina is not of uniform spatial resolution. The density of photoreceptor cells is higher at the center, which enables vision to be more accurate at the center (i.e. the fovea) than at the periphery. There are two types of photoreceptor cells:

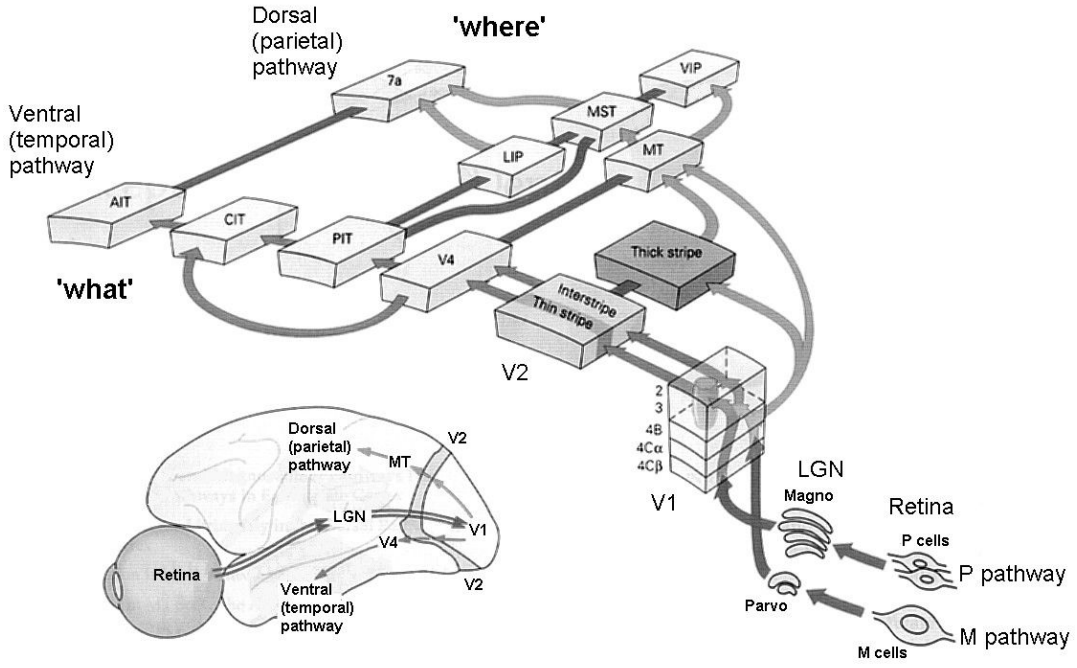


Figure 2.2.1: The human visual system: from retina to different areas of visual cortex (figure from [Perreira Da Silva 10]).

rods and cones, which are sensitive to light and color, respectively. Cone photoreceptors can be divided into three classes based on their spectral sensitivity: "Blue" or short wavelength (S); "Green" or medium wavelength (M); and "Red" or long wavelength (L).

The photoceptor cells are connected to Ganglion cells, which provide the output signal from the retina. The receptive field of ganglion cell is circular and separated into two areas: a center area and a surround area. Two types of ganglion cells exist: the on-center cells which respond excitatorily to light at the center and off-center cells which respond inhibitorily to light at the center [Frintrop 06]. The center area and the surround area always have opposite characteristics. This serves as the reason to consider the mechanism for processing visual information as a center-surround filtering.

Visual Pathway

From the retina the optic nerve transmits visual information to the optic chiasm. From the optic chiasm, two visual pathways go to different areas of brain. In primates, the majority (approximatly 90%) of the visual information is transmitted by the retino-geniculate pathway to the Lateral Geniculate Nuclei (LGN); the remaining 10% goes to the superior colliculi. LGN cell fibers then transmit visual information to the cortical receiving area for vision, known as primary visual cortex or V1, which is located at the very back of the brain. From the primary visual cortex, the visual information is sent

to higher brain areas, namely extrastriate cortex. The extrastriate cortex includes: V2, V3, V4, the infero-temporal cortex, the middle temporal area and the posterior-parietal cortex [Frintrop 06].

There is evidence that the connections between extrastriate areas segregate into two pathways after area V2: a ventral path way and a dorsal path way. The dorsal pathway, which concerns the motion and depth information, runs via V3 to the middle temporal area (MT), then to the medial superior temporal area (MST) and the parieto occipale area (PO) and finally to the posterior-parietal cortex (PP). The dorsal pathway is also named as the "where pathway", since it mainly concern with the question of "where" something is in a scene. The ventral pathway, which concerns the color and form information, runs to V4 and finally in infero-temporal cortex (IT). Since the area IT responds to the recognition of objects, this pathway is also named as the "what pathway".

Attentional mechanism in the brain

So far, it is believed that visual attention is not guided by any single brain area. Several areas have been found to be involved in the attentional process, but the accurate task and behavior of each area, as well as the interplay among these areas, still remain as open questions [Frintrop 06]. Nevertheless, several findings have been claimed. It was proposed that the posterior-parietal cortex responds to disengaging the focus of attention from its present location (inhibition of return); the superior colliculus (SC) is responding for shifting the attention to a new location [Posner 90]. The Frontal Eye Field area of the prefrontal cortex is found to be involved in guiding the eye movements. Additionally, this area is also the place where a kind of saliency map is located, which is affected by both bottom-up and top-down information [Bichot 01].

2.3 Eye movements and eye-tracking

Eye tracking is a technique which records the eye movements so that the researchers can obtain precise information about (1) where a subject is looking at any given time and (2) the sequence in which his eyes are shifting from one location to another.

Eye tracking plays a substantial role in the research of psychology, biology, computer vision, and especially the computational modeling of visual attention. Given the strong link between overt visual attention and eye movements [Itti 01, Wolfe 04], eye movements data collected by means of eye tracking experiment are used as the ground truth to evaluate the performance of computational models.

2.3.1 Measures of eye-movements

Just et al. [Just 76] assumed that what a person is looking at indicates what is at the "top of the stack" in cognitive processes. This "eye-mind" hypothesis implies that

the eye movements provide a trace about where a person's (overt) attention is being directed. There exist various types of eye movements. Two basic ones are "fixation" and "saccade". From these two basic eye movements, another measurement, "scanpath", is stemmed. Moreover, pupil size and blink rate are also two types of eye movements usually studied. Introduction of each type of eye movement as well as metrics based on these basic types of eye movement are presented below.

Fixation

Fixation means that the visual gaze is relatively stationarily maintained on a single location. Fixations last for 218 milliseconds on average, with a range of 66 to 416 milliseconds [Poole 06]. Based on fixation, several metrics can be derived:

- Fixations per area of interest. Experiments show that more fixations on a particular area indicate a greater interest or importance of a target [Wang 10]. And it may also mean that the target is complex in some way and difficult to encode [Just 76]. Jacob et al. [Jacob 03] suggest that, in a search task, a higher number of fixations often means a greater uncertainty in recognizing a target item.

- Fixations duration. A longer fixation can be interpreted in two ways: information is difficult to extract, or the object is more engaging in some way [Just 76].

- Fixation spatial distribution. Cowen et al. [Cowen 02] suggested that highly concentrated fixations in a small area mean a focused and efficient searching, and evenly spread fixations indicate a widespread and inefficient searching. It was also found that if an object contains an area with highly concentrated fixations, the object is tended to be considered as of high importance [Wang 10].

- Repeat fixations, which is also named as "post-target fixations". A higher number of off-target fixations after the target has been fixated (i.e., a lower number of repeat fixations) means that the target lacks meaningfulness or visibility [Goldberg 99].

- Time to first fixation on-target. A shorter time to first-fixation on an object or area indicates that the object or area has better attention-getting properties [Byrne 99].

Note that in the studies of computational modeling of visual attention, fixation spatial density is the metric mostly used, by means of computing a so-called "fixation density map".

Saccades

Saccades are those quick, simultaneous movements of both eyes in the same direction[Cassin 90]. They are fast movements of eyes occurring between fixations. It is generally believed

that no encoding takes place in human visual system during saccades, so vision is suppressed and it is difficult for us to get any clues about the complexity or salience of an object from the saccades. However, information about visual perception can be still extracted from several saccade metrics:

- Number of saccades. A larger number of saccades indicates that more searching takes place during the observation [Goldberg 99].
- Saccade amplitude. Saccade amplitude is computed by measuring the distance between one saccade's start point (a fixation) and its end point (another fixation). Larger amplitude indicates the existence of more meaningful cues, since the attention is drawn from a distance [Goldberg 02].

Scanpaths

Scanpath is a metric derived from the measurement of both fixations and saccades. A scanpath means a complete saccade-fixate-saccade sequence. The area covered by scanpath indicates the area observed. A longer scanpath means a less efficient searching [Goldberg 02]. Additionally, we can compare the time spent for searching (saccades) to the time spent for processing (fixation) in a scanpath. A higher saccade/fixation ratio means more searching or less processing.

Blink rate and pupil size

The blinking of eye and the variation of pupil size are two eye movements that could also be recorded during eye tracking experiments. They can be considered as a cue that indicates cognitive workload. A lower blink rate is assumed to indicate a higher cognitive workload [Bruneau 02], and a higher blink rate may indicate visual fatigue [Brookings 96]. The changing of pupil size also indicates some kinds of cognitive effort [Marshall 00]. However, the blink rate and the pupil size can be easily affected by many factors during the observation, e.g. the luminance of environment. Due to this reason, blink rate and pupil size are not widely used in the researches of visual attention.

2.3.2 Eye-tracking

Eye tracking is a technique which records eye movements so that the researchers can obtain precise information about (1) where a subject is looking at any given time and (2) the sequence in which his eyes are shifting from one location to another. Eye tracking has thus been deployed in a variety of disciplines to capture and analyze overt visual attention of human observers, including neuroscience, psychology, medicine, human factors, marketing, and computer science [Duchowski 02].

The common goal amongst all these disciplines is to capture human viewing behavior when performing specific visual tasks in a given context. For instance, in marketing

research it is of interest to determine what products customers attend to in order to maximize profit [Wedel 07]. In medical research it is of interest to identify the search patterns of radiologists when investigating mammograms for improved breast cancer detection [**?**]. In image and video quality assessment, taking into account the attention of viewers to artifacts may lead to enhanced quality prediction models [Engelke 11]. In the context of computational modeling of visual attention, the eye-tracking results are usually post-processed into scanpaths or so-called fixation density maps (FDM), which are considered to be a reliable ground truth for developing computational models of visual attention.

2.3.2.1 Background of eye tracking technique

The technology of eye tracking appeared firstly more than 100 years ago in reading research [Rayner 89]. Since then, different techniques have been applied in eye tracking. For instance, the "electro-oculographic techniques" needs to put electrodes on the skin around the eye so that eye movements can be detected by measuring the differences in electric potential. Some other methods relied on wearing large contact lenses. The lenses covered the cornea (the transparent front part of the eye) and sclera (the white part of the eye), while a metal coil was embedded around the lens so it moved along with the eye. The eye movements could be thus measured by fluctuations in an electromagnetic field when the eye was moving [Duchowski 07]. However, these historical methods affect observers' eye-movement and are inconvenient to implement.

Video-based techniques are used by the modern eye-trackers to determine where a person is looking (i.e., the so-called "gaze point" or "point-of-regard"). These eye-trackers achieve the detection of point-of-regard based on the eye's features extracted from video images of the eye, such as corneal reflections (i.e. Purkinje images), iris-sclera boundary, and the apparent pupil shape [Poole 06, Duchowski 07].

Most state-of-the-art commercial eye trackers use the "corneal-reflection/pupil-centre" method to measure the point-of-regard. The corneal reflection is also known as (first) Purkinje image. During the eye tracking, a camera focuses on one or both eyes to get images. Contrast is then used to get the location of the pupil, and infrared light is used to create a corneal reflection. By measuring the movements of corneal reflection relative to the pupil, it is then possible to know the head movement, eye rotation, the direction of gaze and consequently the point-of-regard.

2.3.2.2 Algorithms for eye-movement identification

Given the information about eye-movement type (e.g. fixations, saccades) and their characteristics (e.g. duration, spatial distribution), various subsequent analyses can be then performed depending on the particular context and application of the research. However, the raw eye-movement data output from eye-tracking experiments are usually presented by means of a stream of sampled gaze points. Post-processings need to be

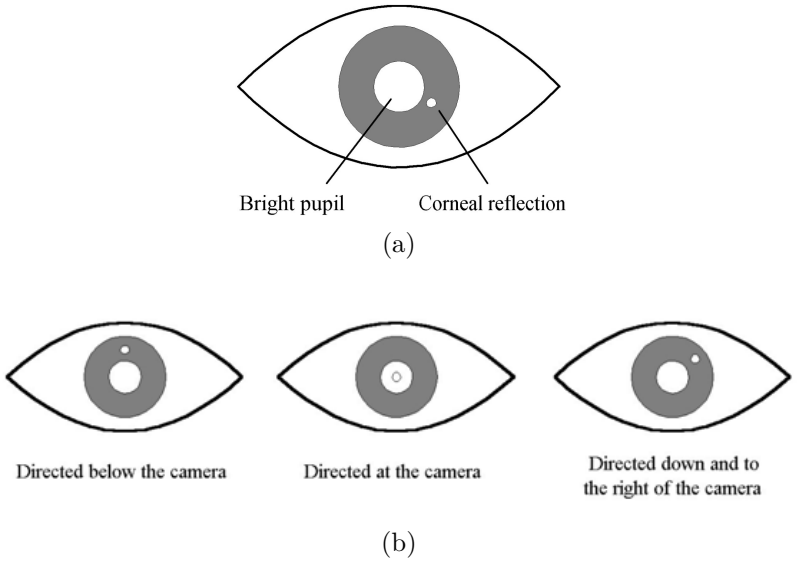


Figure 2.3.1: (a) Illustration of corneal reflection and pupil seen in an infrared camera image. (b) Illustration of how the corneal reflection position changes according to the point-of-regard [Poole 06].

firstly performed to identify different types of eye movements from the gaze points.

The fixation detection algorithms extract and label fixations and saccades from raw eye-tracking data (i.e. sample points). These algorithms can identify the fixations, the saccades taking place between two successive fixations, and those smaller eye movements occurring during fixations, such as tremors, drifts, and flicks[Salvucci 00]. Note that the fixation identification is a critical aspect of eye-movement data analysis, since its result can significantly affect later analyses. Evidences have showed that different identification algorithms could produce different interpretations even when analyzing the same eye-tracking data [Karsh 83].

Salvucci et al. [Salvucci 00] suggested that most fixation identification algorithms took advantage of the following spatial or temporal features:

- Velocity. Some algorithms take advantage of the fact that fixation points have much lower velocities compared to the saccades. Generally, the sampling rate of an eye-tracker is constant, so the velocity equals to the distance between sample points.

- Dispersion. Some algorithms emphasize the spread distance (i.e. dispersion) of fixation points. It assumes that the sample points belonging to a fixation generally occur near one another, but saccades are far away from others.

- Duration information. This criterion is based on the fact that fixations are rarely less than $100\,ms$ and usually in the range of $200 - 400\,ms$.

- Local adaptivity. This criterion means that the interpretation of a given point is influenced by the interpretation of temporally adjacent points.

Based on the different features selected, various fixation identification algorithms have been proposed. Two principle types of fixation identification algorithms are introduced below.

Velocity-based Algorithms

The velocity information of eye movements shows two distributions of velocities: low velocities for fixations, and high velocities for saccades. These velocity-based discrimination is straightforward and robust.

Among various velocity-based algorithms, Velocity-Threshold Identification (I-VT) is the simplest one to implement [Salvucci 00]. I-VT calculates firstly point-to-point velocities for each point. Each velocity is computed as the distance between the current point and the next (or previous) point. Each point is then classified as a saccade point or fixation point based on a velocity threshold: if the velocity is higher than the threshold, it becomes a saccade, otherwise it becomes a fixation point. Finally, I-VT translate each fixation group into a <x, y, t, d> representation. <x, y> represent the centroid of the points, t and d means the time of the first point and the duration of the points respectively.

A more sophisticated type of velocity-based algorithm is Hidden Markov Model fixation Identification (I-HMM) [Salvucci 99, Salvucci 98]. I-HMM applies a two-state HMM in which the two states represent the velocity distributions for saccade and fixation points, respectively. Generally, I-HMM can perform more robust identification than fixed-threshold methods (e.g. I-VT) [Salvucci 00].

Dispersion-based Algorithms

Dispersion-based Algorithms utilizes the fact that fixation points tends to cluster closely together because of their low velocity. Dispersion-Threshold Identification (I-DT) is a typical type of the dispersion-based algorithms. I-DT identifies fixations as groups of consecutive points within a particular dispersion. A dispersion threshold is thus essential for I-DT algorithms. Moreover, a minimum duration threshold is also required, which is used to help alleviate equipment variability. The minimum duration threshold normally ranges from 100 ms to 200 ms [Widdel 84].

An implementation of I-DT algorithm is proposed by Widdel et al. [Widdel 84]. They use a moving window to cover consecutive data points. The moving window begins at the start of the protocol. It contains initially a minimum number of points which is determined by a given duration threshold. The I-DT then compute the dispersion of the points in the window by summing the differences between the points' maximum and minimum x and y: $D = [max(x)–min(x)] + [max(y)–min(y)]$. To help alleviate equipment variability, it incorporates a minimum duration threshold 100-200ms [Widdel 84].

If the dispersion is above a dispersion threshold, the window move to the following point. If the dispersion is below the threshold, the window represents a fixation and will be expended until the window's dispersion is above the threshold. The final window is marked as a fixation which centers at the centroid of the points and has a given onset time and duration.

2.4 State-of-the-art in computational modeling of visual attention

Eye-tracking experiment can be considered as a reliable way to acquire the distribution of human's attention on a specific scene. However, conducting eye-tracking experiments is usually cumbersome, time consuming, and hence, expensive. In order to automatically predict the distribution of human's attention, extensive research efforts have been dedicated in computational modeling of visual attention. In our study, we particularly focus on the models that compute saliency maps. The results of this type of model, the saliency maps, indicate where the most visually interesting regions are located.

In the past years, a body of models using various mathematical tools have been proposed. According to the taxonomy introduced by Le Meur and Le Callet [Le Meur 09], most of the computational models can be grouped into three main categories: hierarchical model, statistical model, and Bayesian model.

2.4.1 Principle computational models

2.4.1.1 Hierarchical model.

The computational architectures of the hierarchical models are similar. This kind of models is characterized by the use of a hierarchical decomposition, whether it involves a Gaussian, a Fourier-based or wavelet decomposition. Various feature maps are then computed. Different strategies are then used to integrate information across subbands to create a final saliency map.

The model of Itti

One of the most famous models of this category is the model proposed by Laurent Itti in 1998 [Itti 98]. It is the first computational and biologically plausible model of bottom-up visual attention, and serves as a basis in many studies. The architecture of this model (see Figure 2.4.1) is based on the following principle steps. The original image is firstly decomposed into three different perceptual channels: intensity, color and orientation. A multi-scale representation is constructed from the various channels. A set of linear center-surround operations akin to visual receptive fields are performed to obtain feature maps. These feature maps are (1) normalized by an operator $\mathcal{N}$ which enhances the

feature maps containing a small number of peaks of saliency, and (2) then summed to create the conspicuity maps (intensity, color, orientation). These three conspicuity map are also normalized (by the operator $\mathcal{N}$) and summed to obtain a unique saliency map. In this model, the selection of focuses of attention (FOA) is achieved by a "winner-take-all" network, which selects the most salient area in the saliency map and contains a "inhibition of return" mechanism to temporarily prevent the FOA return immediately to the areas already visited.

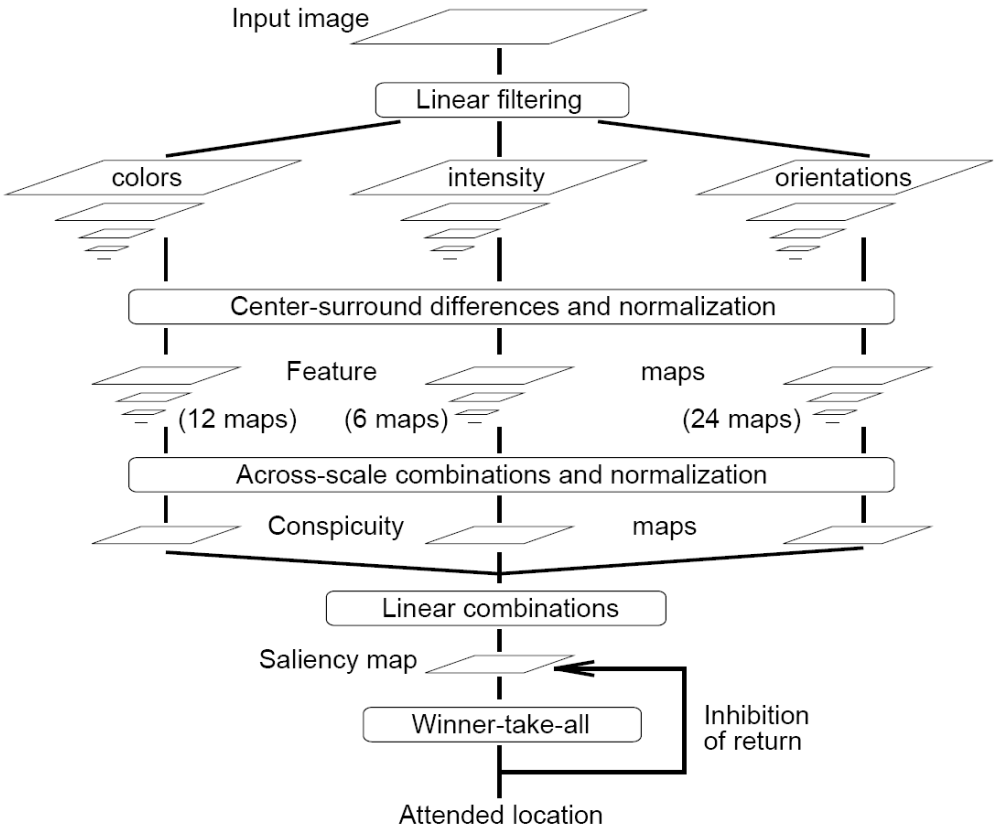


Figure 2.4.1: Architecture of the Itti's model [Itti 98].

The model of Le Meur

Another representative model comes from Le Meur et al. [Le Meur 06]. It is also a bottom-up model based on Treisman's Feature Integration Theory [Treisman 80] and the biologically plausible architecture proposed by Koch and Ullman [Koch 85]. This model was first described in [Le Meur 06] and then modified in [Le Meur 07], which takes into account movement. What we introduce here is the original version of the model.

Le Meur's model (see Figure 2.4.2) builds on a coherent psychovisual space. Three aspects of the vision process are tackled: the visibility, the perception, and the perceptual grouping. The "visibility" process simulates the limited sensitivity of the human visual system. For an input image, RGB luminance is first transformed into the Krauskopf's color space (A, C_{r_1} and C_{r_2}), which simulates the three channels used by retina for visual information encoding. The first channel, A, transforms achromatic perceptual signals; the second channel, C_{r_1}, transforms chromatic perceptual signals of the opponent colors of red-green; and the third channel, C_{r_2}, transforms chromatic perceptual signals of the opponent colors of blue-yellow. A contrast sensitivity function is then applied to each of the three channels. These contrast sensitivity functions show how the sensitivity of human eye varies as a function of spatial frequency and orientation. A hierarchical decomposition is then applied to each of the three channels. The decomposition consists of splitting the 2D spatial frequency domain both in spatial radial frequency and in orientation. In this model, each channel is considered as the feature map corresponding to a specific set of neurons. A "perception" process is then applied, in which a center-surround mechanism is performed to simulate the HVS for selecting relevant areas and reducing the redundant incoming visual information. The third process of the model is "perceptual grouping". It refers to the human visual ability which groups and binds visual features to organize a meaningful higher-level structure. Finally, this computational model sums the output of the different channels to obtain a two-dimensional spatial saliency map. Note that in [Le Meur 07], Le Meur et al. proposed a computational model for video, in which movement is considered as a additional visual channel. The result of the movement channel is a temporal saliency map, which is finally combined with the spatial saliency map to get a final saliency map.

2.4.1.2 Statistical model.

This kind of models utilizes probabilistic methods to compute the saliency. The probabilistic framework is deduced from the content of current image. The measure of saliency of each location is based on various features, and is defined as the deviation of these features between the current location and its surrounding region. Note that (1) the features used in statistical models refer not only to the low level visual features (e.g. color, intensity or orientation) but also some features derived by Independent Component Analysis (ICA) or Principal Component Analysis (PCA) algorithms; and (2) even the whole image might be considered as the "surrounding region" in these models.

The model of Bruce and Tsotsos

In 2009, Bruce et al. [Bruce 09] proposed a model of saliency computation based on the premise that localized saliency computation serves to maximize information sampled from one's environment. The framework of this model is depicted in Figure 2.4.3.

The first step of this model is the independent feature extraction. For each location

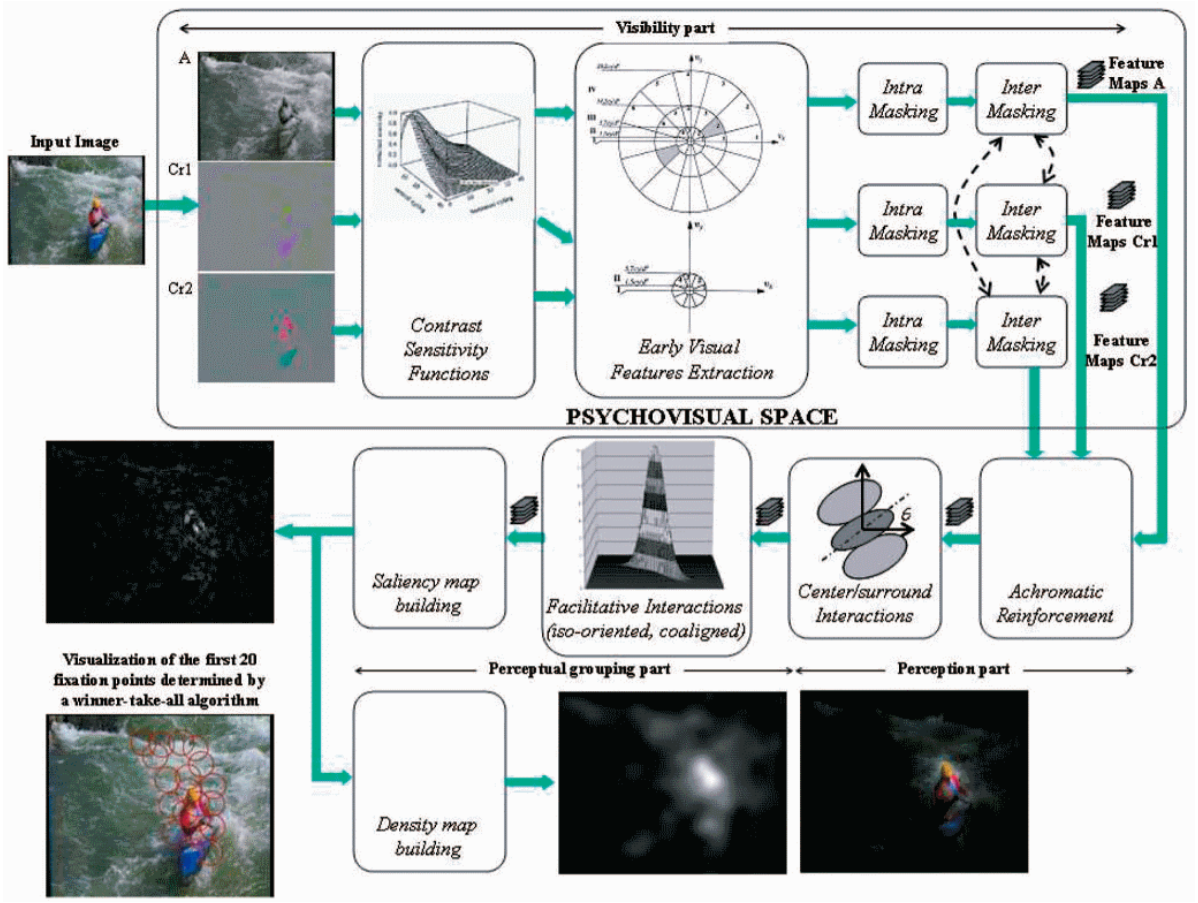


Figure 2.4.2: Architecture of the Le Meur's model [Le Meur 06].

(i, j) in the image, the response of various learned filters that simulate V1 cortical cells are computed. This operation is considered as measuring the response of various cortical cells coding for content at each individual spatial location. Gabor-like cells that respond to orientation structure within a specific spatial frequency band and cells that respond to color opponency are taken into account. This step yields a group of coefficients for each local neighborhood of the scene $C_{i,j}$.

The second stage is density estimation. The content of each local neighborhood $C_{i,j}$ of the image is characterized by several coefficients a_k. These coefficients, a_k, correspond to the various basis filters coding for that location. At one spatial location and in the surrounding regions of that location, there are a set of coefficients for a same filter type. Based on a non-parametric or histogram density estimate, the coefficients in the surround form a distribution that can be used to predict the likelihood of the coefficients of $C_{i,j}$. Any given coefficient can be then converted to a probability by looking up its likelihood from the probability distribution derived from the surround. Based on the probabilities, joint likelihood of each location can be computed, which is then translated into Shannon's measure of Self-information. The resulting information map serves as the output of the model, the spatial saliency map.

The model of Gao et al.

The model proposed by Gao et al. [Gao 08] computes a so-called "discriminant center-surround saliency" by combining (1) the classical assumption that bottom-up saliency is a center-surround process, and (2) a discriminant saliency hypothesis. The computation of saliency is formulated as a binary classification problem (see Figure 2.4.4). For each location l in the input image, the saliency is defined with respect to two classes of stimuli: stimuli of interest and null hypothesis. Stimuli of interest refer to the observations within a neighborhood of l, (i.e. W_l^1, which is referred to as the center); null hypothesis refers to the observations within a surrounding window, (i.e. W_l^0, which is referred to as the surround). The saliency of each location is thus equal to the discriminant power, which is quantified by mutual information, for the classification of the observed features that comes from the center area and the surround.

2.4.1.3 Bayesian model.

In addition of information from the current image, Bayesian framework is applied in this type of models to take into account also the prior knowledge. This prior knowledge concerns, for instance, the statistic of visual features in natural scenes, including its distribution or its spectral signature. Since prior knowledge from the perceptual learning process would help the human visual system to understand the visual environment, the integration of prior knowledge into computational model could be compared to a visual priming effect that would facilitate the scene perception [Le Meur 09].

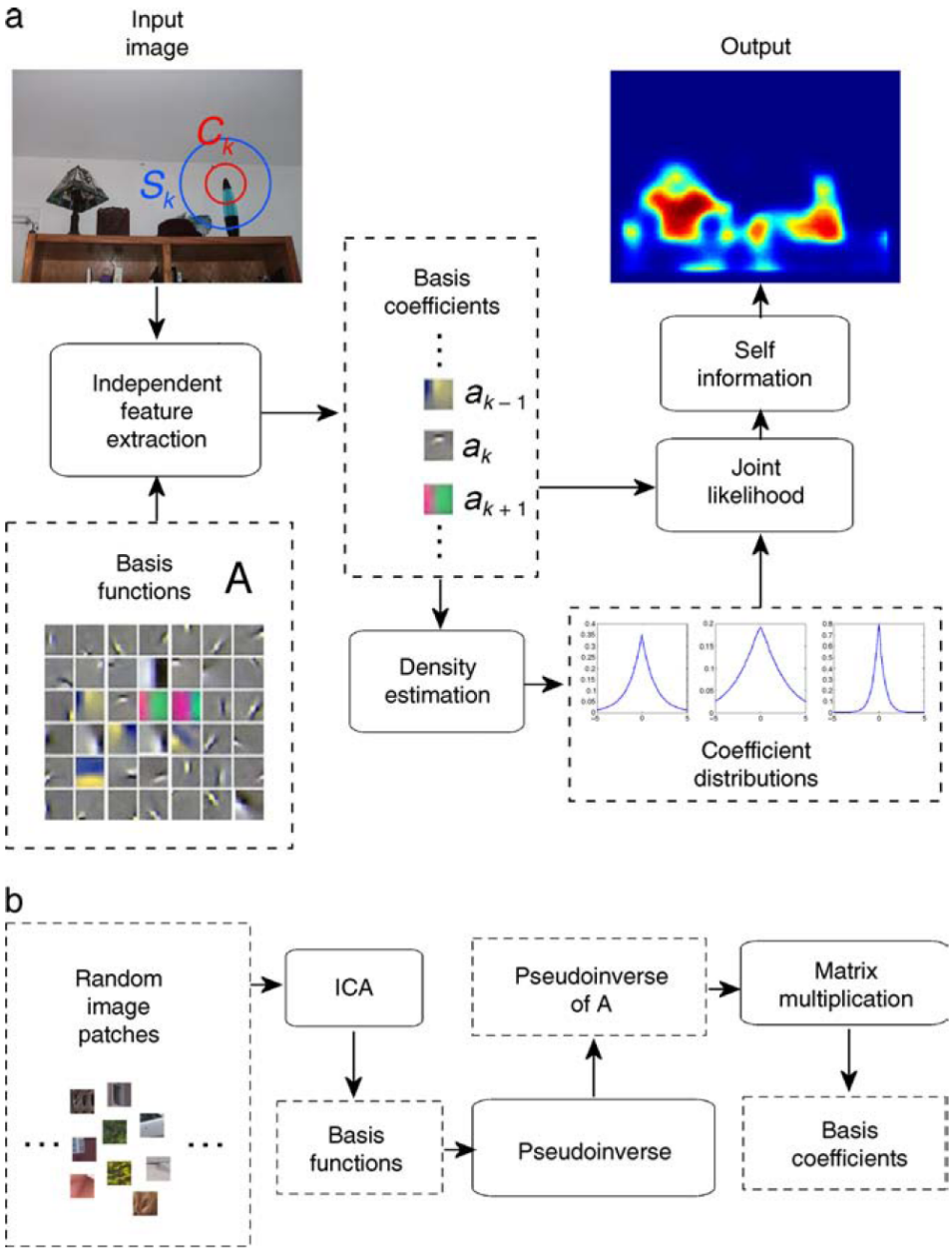


Figure 2.4.3: Architecture of model of Bruce and Tsotsos [Bruce 09].

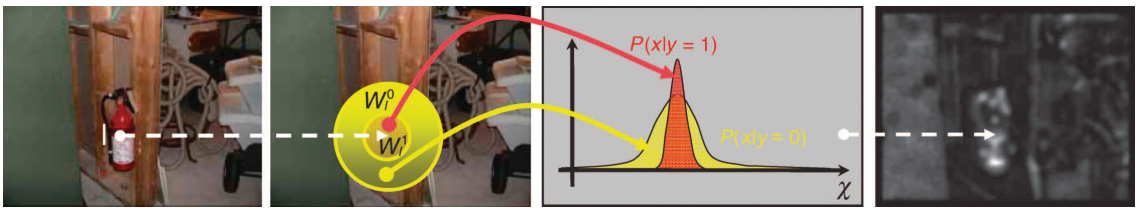


Figure 2.4.4: Computation of the "discriminant center-surround saliency" [Gao 08].

The model of Zhang et al.

The model proposed by Zhang et al. [Zhang 08] is based on an assumption that one goal of human visual system is to find potential targets by estimating the probability of a target at every location given the visual features. The proposed model relies on a Bayesian probabilistic framework, in which bottom-up saliency is regarded as the self-information of visual features; when searching for a target, the overall saliency is considered as the point-wise mutual information between the features and the target. By (1) letting the binary variable C denote whether a point belongs to a target class, (2) letting the random variable L denotes the location, (3) letting the random variable F denote the visual features, the computation of saliency of specific location z (e.g. a pixel) is formulated by:

$$S_z = p(C = 1|F = f_z, L = l_z)$$

where f_z represents the feature observed at z, and l represents the location (i.e. pixel coordinates) of z.

Compared to other bottom-up saliency measures, which are defined solely in terms of the image currently being viewed, this model is defined based on natural statistics collected from a set of images of natural scenes. And this is the reason why it is named SUN. Besides, compared to the others, it involves only local computation on images, without calculation of global image statistics or saliency normalization or winner-take-all competition.

Within the Bayesian framework proposed, z denotes a point (in this model, it is a pixel of the image), C denotes whether or not a point belongs to a target class, L denotes the location, F denotes the visual features of a point. Saliency of z can be defined as $p(C = 1|F = f_z, L = l_z)$. Here, f_z represents the feature values observed at z, l_z represents the location of z. Bayes' rule can be used here to calculate this probability:

Due to the assumptions that (1) features and location are independent and conditionally independent given C = 1, and (2) the distribution of a feature does not change with location, the formulation is given by:

$$\log S_z = -\log p(F = f_z) + \log p(F = f_z, C = 1) + \log p(F = f_z, L = l_z)$$

The first term on the right side of this equation, $-\log p(F = f_z)$, is the self-information. The rarer the visual features are, the more informative they are. The second term,

$\log p(F = f_z, C = 1)$, is a log-likelihood term which favors feature values consistent with our knowledge of the target. It corresponds to the top-down effect when searching for a known target. The third term in the equation, $\log p(F = f_z, L = l_z)$, is independent of visual features and represents any prior knowledge of where the target is likely to appear. In the free-viewing condition, both the the location prior knowledge and the log-likelihood term are unknown, so the bottom-up saliency is equal to the self-information, $-\log p(F = f_z)$.

2.4.2 Features for visual saliency detection

The selection of visual features is of great importance in the computational modeling of visual attention. According to the feature integration theory [Treisman 80], three features have been widely used in existing computational models of visual attention: intensity, color and orientation [Borji 12]. Intensity is usually processed by a center-surround process, which is inspired by neural responses in lateral geniculate nucleus (LGN) and V1 cortex. To extract this feature, two types of filter are used to simulate the response of visual cells that have a center ON (resp. OFF) and a surround OFF (resp. ON). Color is usually taken into account by means of the red/green and the blue/yellow color pairs, which is inspired by color-opponent neurons in V1 cortex. Orientation is usually implemented as a convolution with oriented Gabor filters or by the application of oriented masks. Movement is also used in the models for video (in the primate brain motion is derived by the neurons at MT an MST regions which are selective to direction of motion [Borji 12]). In addition to the basic visual features introduced previously, some other specific features that direct human's attention have been used in the modeling of visual attention [Borji 12], including: face [Cerf 08], horizontal line [Oliva 01], wavelet [Li 10], gist [Torralba 03], center-bias [Tatler 07], spatial resolution [Hamker 05], optical flow [Vijayakumar 01], flicker [Itti 03], crosses or corners [Privitera 00], entropy [Kadir 01], ellipses [Lee 05], symmetry [Kootstra 08], texture contrast [Parkhurst 02], depth [Maki 00], components derived by ICA or PCA algorithms [Zhang 08, Bruce 09].

2.5 Conclusion

In this chapter, we firstly introduce the concept of visual attention as well as various mechanisms of attention, for instance, the distinction of overt/covert attention and bottom-up/top-down attention. Secondly, we present a brief introduction of the HVS. In this part, instead of presenting an exhaustive explanation of the whole HVS, we particularly focus on introducing the retina and different areas of the visual cortex, both of which allow to determine the main characteristics of the HVS. Next, we introduce different types of eye movements as well as eye-tracking. State-of-the-art techniques of measuring gaze points and fixation identification algorithms are introduced. Finally, we introduce some principles and famous state-of-the-art computational models of visual

attention.

Being the first chapter of the whole thesis, this chapter provides a review of the background and state-of-the-art studies that are important in the research field of visual attention. One can find that the research on visual attention is a highly interdisciplinary field; different relevant disciplines deal with the research on visual attention from different points of view, and profit from each other. In the following chapters, we will present our (experimental and modeling) studies in the filed of visual attention.

Key points

Context

- ❏ Visual attention is a key mechanism in the HVS to reduce the large amount of information received, which is beyond the capability of HVS to deal with all of them.
- ❏ Based on different attributes, attention can be differentiated in different ways. Based on the relationship between attention and eye movements, attention is classified into overt attention and covert attention; based on the different factors that shift the attention, attention is classified into bottom-up attention and top-down attention
- ❏ Due to the link between (overt) attention and eye movements, eye-tracking has been a technique that has been widely used to study visual attention.
- ❏ Based on different computational architectures and visual features, a large amount of computational models of visual attention have been proposed. The output of these models, saliency map, indicates where the most visually interesting regions are located.

Contributions

- ❏ In this chapter, we provide a review of the background and state-of-the-art studies that are important in the research field of visual attention.

Clarifying and designing visual attention related ground truths for image processing applications

"Although nature commences with reason and ends in experience it is necessary for us to do the opposite, that is to commence with experience and from this to proceed to investigate the reason."

(Leonardo da Vinci)

Chapter 3

Comparative study of fixation density maps obtained from different experimental conditions

Fixation density maps (FDM) created from eye tracking experiments have been widely used in both studies of visual attention and image processing applications. Since the FDM are assumed to be reliable ground truths about the attentional behavior of human observers, the accuracy of FDM is of particular importance. Nowadays, the community already has many FDM databases provided by different laboratories. These FDM are similar, and they are all expected to provide the ground truth for the image processing community. However, no studies so far have analyzed the difference between them and the related impact on the applications.

Since this is the "first" chapter (not including the introduction chapters) of a thesis regarding visual attention, we start our work with a study focusing on the accuracy of ground truth, i.e. FDM. In this chapter, we perform a thorough comparison of three FDM databases, which are created by 3 different laboratories[1]. We focus on the effect of presentation time and image content, and we evaluate the impact of the FDM difference on three applications: visual saliency modeling, image quality assessment, and image retargeting.

The remainder of this chapter is organized as follows. Section 3.1 presents an introduction of the background of this study. Section 3.2 introduces the eye tracking experiments and discusses the post-processing of the gaze patterns. The similarity measures that are used throughout the article are presented in Section 3.3. Section 3.4 then provides a detailed comparison of the FDM between the three conducted experiments. The influence of the FDM on three applications is discussed in Section 3.5 and a detailed discussion

[1]This study is performed though an international collaboration with Philips Research Laboratories (The Netherlands), Delft University of Technology (The Netherlands), Blekinge Institute of Technology (Sweden) and University of Western Sydney (Australia).

is provided in Section 3.6. Finally, conclusions are drawn and an outlook is given in Section 3.7.

3.1 Context

Given the strong link between overt visual attention and eye movements, the fixation density map (FDM) obtained by eye-tracking experiment is believed to provide reliable ground truth of overt visual attention. However, eye-tracking experiments are usually cumbersome, time consuming, and hence, expensive. For this reason, there has been a strong demand for publicly available eye-tracking databases. In the context of image and video processing applications, such databases also facilitate fair comparisons amongst computational models. This need has been realized by both the computer science as well as the vision science community. For this reason, several image and video eye-tracking databases have been made publicly available in recent years [Bruce 09, Wang 10, Judd 09, Engelke 09a, Liu 09, Cerf 09, Itti 98].

So far, there are no standardized methodologies for the conduction of eye tracking experiments. Instead, researchers usually follow best-practice guidelines [Duchowski 02]. The outcomes of the experiments hence depend on different factors related to the observer and the experimental design. The observers, for instance, differ with respect to their cultural background, age, gender, experience, and expectations. It is known that humans have varying interest in visual content, hence the resulting FDM can vary considerably between participants [Judd 10]. These variations are the main reason why averaged FDM instead of individual FDM are used for VA model design. Apart from the observer variations, also environmental aspects related to the procedures of the eye tracking experiment may affect the final FDM (e.g. the eye tracker hardware, the display, the viewing conditions, and the stimuli presentation). The lack of agreement on these experimental procedures can lead to considerable differences in the resulting FDM, even if they are averaged over a large observer population. Moreover, the algorithm applied for identifying eye movements (i.e. fixations) has great impact on the final FDM.

To identify the reliability of the FDM as a ground truth, it is crucial to evaluate the similarity of FDM created from experiments that were conducted in different laboratories. This is of particular importance if the experiments were performed independently, in which case many design parameters may be different. Inter-laboratory comparisons are common in research disciplines related to natural sciences [Doležel 98] and medical sciences [Davey 94], however, such comparisons are less common in computer science society. In computer science society, researchers often restrict themselves to assess the differences amongst observers within an experiment. This has been previously studied for the context of eye tracking on images [Engelke 10] and on video [Dorr 10].

To the best of our knowledge, there has not been any comprehensive study on the comparison of eye-tracking data amongst different laboratories. In this chapter, we therefore study the degree to which FDM of images differ between three experiments. These ex-

periments were not conducted conjointly for the purpose of comparing experimental FDM, while they were performed independently with each experiment considering their FDM to be solid ground truths. The goal here is therefore not to compare FDM that were created using exactly the same setup, but rather to analyze the differences amongst FDM and to estimate their reliability as a ground truth.

We particularly focus on the influence of two factors: visual content and image presentation time. The former factor can be assumed to have a strong impact. It is expected that the agreement between observers varies with respect to the degree to which the objects in the scene attract the viewers' attention. The latter factor, presentation time, is of interest for two reasons: (1) the similarity between FDM of different laboratories may vary according to the duration for which the images are presented; and (2) the FDM may experience convergent behavior, meaning that after a certain presentation time the changes of the FDM are becoming negligible. Hence, identifying the time dependency of FDM aids in determining presentation times that are sufficient to obtain convergent FDM. This result can help to reduce experimental time and cost. Moreover, we address also the impact of FDM similarity on three applications: visual saliency modeling, computational quality assessment, and image retargeting. Finally, we discuss the influence of different subjective and environmental factors that impact on the similarity between the FDM.

3.2 Eye-tracking experiments

To perform a comprehensive study on the comparison of eye-tracking data collected from different laboratories, we conducted an eye-tracking experiment. Additionally, two other eye-tracking experiments were conducted in two independent laboratories at (1) the School of Computing and Mathematics at the University of Western Sydney (UWS), Australia [Engelke 09a], and (2) the Man-Machine Interaction group of Delft University of Technology (TUD), The Netherlands [Liu 09]. These three experiments are in the remainder of this chapter referred to as UN, UWS, and TUD, respectively.

3.2.1 Stimuli

The images that were presented in all the three experiments came from the LIVE image quality database [Sheikh 05]. This database contains a total of 779 distorted images based on 29 original (reference) images. For the eye-tracking experiments, only the 29 reference images were used. These images cover a wide range of content, including natural scenes, buildings, boats, humans, animals, and written text. Therefore, the use of these images facilitates to identify the impact of different contents on the variations of inter-laboratory eye-tracking experiments. Table 3.1 gives a list of all the images with the original names from the LIVE database. The corresponding numbers are used throughout the rest of this chapter to associate the analysis results with the images.

#	Name	#	Name	#	Name
1	bikes	11	house	21	saling1
2	building2	12	lighthouse	22	saling2
3	buildings	13	lighthouse2	23	saling3
4	caps	14	manfishing	24	saling4
5	carnivaldolls	15	monarch	25	statue
6	cemetry	16	ocean	26	stream
7	churchandcapitol	17	paintedhouse	27	studentsculpture
5	coinsinfountain	18	parrots	28	woman
9	dancers	19	plane	29	womanhat
10	flowersonih35	20	rapids		

Table 3.1: Reference images in the LIVE image quality database [Sheikh 05].

3.2.2 Comparison of experimental procedures

An overview of the three experiments is presented in Table 3.2. Additionally, some differences, which might be a source of variability in the recorded data, are introduced as follow.

Observer

In terms of the difference (between the experiments) regarding participants, firstly, a major difference is the number of observers. In the three experiments, the number of observers ranged from 15 (for UWS) to 21 (for UN). Generally, eye-tracking data averaged over a group of observers becomes more stable with an increased number of observers [Duchowski 02]. Secondly, the average age between UWS and UN is considerably different and may also have an impact on the viewing behavior, since people of different ages have different interests. No age was recorded in experiment TUD. However, given that all participants were students, the average age is expected to be the lowest among the three experiments. Finally, as the experiments were conducted in three different countries (i.e. France, Australia, and The Netherlands), cultural differences between the observers might also have an influence on the resulting eye-tracking data.

Eye-tracker

Three different eye-trackers were used in the experiments (see Figure 3.2.1). In experiment UN, the considerably higher frequency (500 Hz) of the eye-tracker is instrumental for the analysis of saccades, while the frequencies of 50 Hz or below (as for the eye-trackers in UWS and TUD) may not be sufficient for that. Nevertheless, since the analysis focuses on fixations, the recording frequency of the eye-trackers is not expected to have a strong impact on the results. In addition to the higher frequency, the fact

Details	UWS	TUD	UN
Number	15	18	21
Age range (average age)	20-60 (42)	-	18-42 (26)
Male/female	9/6	11/7	11/10
Non-experts/experts	12/3	18/0	21/0
Occupation	University staff/students	University students	University staff/students
Compensated	No	No	Yes
Environment	Laboratory		
Illumination	Low		
Viewing distance	60 cm	70 cm	70 cm
Task	Free-viewing: the observers were not instructed with any particular task but to view the images		
Make	Samsung SyncMaster	iiyama	DELL
Type	LCD	CRT	LCD
Size	19”		
Resolution [pixels]	1280×1024	1024×768	1280×1024
Make	EyeTech TM3 [27]	SMI iView X RED [28]	SMI iView X Hi-Speed [29]
Type	Infrared video-based		
Frequency	45 GP/s	50 GP/s	500 GP/s
Accuracy	$<$ 1 dva	0.5-1 dva	0.25-0.5 dva
Mounting	Under the display	Under the display	Tower with head rest
Calibration	16 point screen	9 point screen	9 point screen
Order	Random		
Image duration	12 s	10 s	15 s
Grey-screen duration	3 s		
Max. visual angle [pixels/deg]	36	32.8	41.8
Central fixation point	Yes	No	No

Table 3.2: Overview of the three eye-tracking experiments.

that the eye-tracker in UN utilizes a tower with a head rest and its comparably higher accuracy can give rise to the assumption that the recorded data is to some degree more accurate at recording gaze positions compared to UWS and TUD.

Viewing condition

Amongst the three experiments, the viewing conditions differed mainly in (1) the viewing distance and (2) the physical image size which is due to different screen resolutions. The visual angle ν, measured in degree of visual angle (dva), accounts simultaneously for these two factors. We computed it here for the maximum image width of 768 pixels in our image set. The horizontal visual angles for UN, UWS, TUD are $\nu_{UN} = 18.38\,dva$, $\nu_{UWS} = 21.33\,dva$, and $\nu_{TUD} = 23.41\,dva$, respectively.

Additionally, the image presentation time differs mainly in two factors. Firstly, the duration ranges from $10\,s$ to $15\,s$. This variation of time means that, for the purpose of comparing the three experiments, we are limited to the first $10\,s$ of each experiment. Secondly, experiment UWS utilized a central fixation point during presenting the gray screen between the presentations of two different images. The central fixation point was used to attract observer's attention, and thus ensured that the viewing of each image was started from the same location (i.e. the center of screen). The other two experiments did not introduce such a center fixation point. This factor is expected to have a considerable impact, especially on the early fixations during image viewing.

(a) (b) (c)

Figure 3.2.1: Eye-trackers: (a) SMI iView X Hi-Speed (UN), (b) EyeTech TM3 (UWS), (c) SMI iView X RED (TUD).

3.2.3 Creation of fixation density maps

In this study, the gaze points obtained by eye tracker are first post-processed into FDM for the purpose of comparison. The FDM (which are also called as "human saliency map") are normalized intensity maps with intensity values ranging from 0 and 1. The magnitudes within the FDM represent the amount of overt attention of the observers. Nevertheless, they do not account for the temporal order of the fixations. Note that fixation order is difficult to predict using a computational model [Perreira Da Silva 10], for which reason FDM are typically used.

The conversion of gaze points into FDM for the three experiments are conducted independently. Despite different softwares were used to create the FDM, the processings were comprised of the same steps:

- Firstly, gaze points belonging to saccades were removed since vision is greatly suppressed during these fast eye movements (as introduced in Section 2.3.1).
- The remaining gaze points were clustered into fixations that determine the locations in the image that the observer focussed on. Note that all the three experiments used velocity-based algorithms for determining the fixations and filtering out the saccades.
- Finally, the fixation map was filtered using a Gaussian kernel to account for (1) the decrease in visual accuracy with increasing eccentricity from the fovea, and (2) the decrease of eye tracker accuracy. All three experiments assumed a minimum fixation length of 100 ms and a foveal coverage of approximately 2 dva.

We created FDM based on a range of presentation times $t \in \{0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ s to analyse the differences between the experiments for different viewing durations. In the

remainder of this chapter, the FDM for image i, created from a particular presentation time t, and belonging to one of the three experiments (i.e. UN, UWS, and TUD), are denoted as $M_{UN}^{(t)}(i)$, $M_{UWS}^{(t)}(i)$, and $M_{TUD}^{(t)}(i)$, respectively. The FDM of all images in the database are presented in Figure 3.2.2 for all three experiments and for a presentation time of $10\,s$.

In addition to the experimental FDM, we created also random FDM for each image. The random FDM represent a "lower limit", both for the similarity evaluation between FDM as well as the performance evaluation of the applications in Section 3.5. The random FDM were created by randomly substituting the FDM between images of the same or similar size within the same database. For better comparability, the same random substitution has been used for all three databases. The random FDM are in the following denoted as $M_{RND}^{(t)}(i)$.

3.3 Similarity measures

So far, there are no standardized measures to compare the similarity between two FDM. However, there is a range of measures that are widely used to perform the comparison between FDM and predicted saliency maps from computational models: correlation coefficient [Le Meur 06, Rajashekar 08], Kullback-Leibler divergence [Le Meur 06, Bruce 09], receiver operating characteristics (ROC) analysis [Zhang 08, Le Meur 10a, Zhao 11], and normalised scanpath saliency (NSS) [Dorr 10, Zhao 11, Le Meur 10a, Marat 09]. The former three are directly applicable to saliency maps and FDM, whereas NSS compares the actual fixations to a saliency map. In this study, we utilise two similarity measures: the Pearson linear correlation coefficient and the area under the ROC curve. Additionally, the FDM of all the images are also provided (see Figure 3.2.2) for qualitative assessments.

3.3.1 Pearson linear correlation coefficient (PLCC)

The Pearson linear correlation coefficient (PLCC) measures the strength and direction of a linear relationship between two variables. We compute it between two FDM, $M^{(i)}$ and $M^{(j)}$, as follows

$$\rho_P(M^{(i)}, M^{(j)}) = \frac{\sum_k \sum_l (M_{kl}^{(i)} - \mu^{(i)})(M_{kl}^{(j)} - \mu^{(j)})}{\sqrt{\sum_k \sum_l (M_{kl}^{(i)} - \mu^{(i)})^2} \sqrt{\sum_k \sum_l (M_{kl}^{(j)} - \mu^{(j)})^2}}$$

where $k \in [1, K]$ and $l \in [1, L]$, are the horizontal and vertical pixel coordinates, respectively; and $\mu^{(i)}$ and $\mu^{(j)}$ are the mean pixel values of the two FDM.

For purposes of illustration, two scatter-like plots are presented (see Figure 3.3.1) for two highly correlated FDM ($M_{UWS}^{(10)}(6)$ and $M_{TUD}^{(10)}(6)$) with $\rho_P = 0.933$) and two lowly

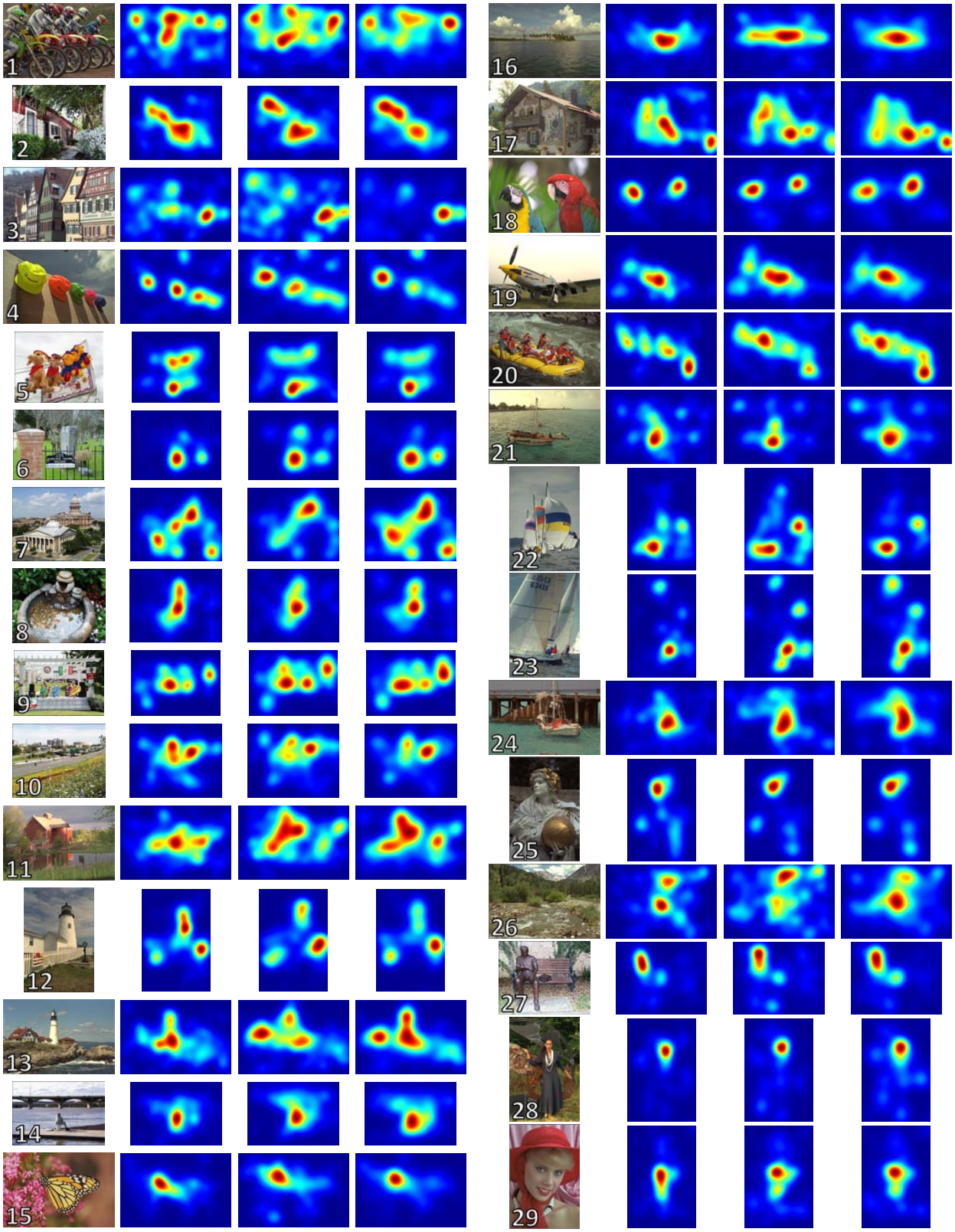
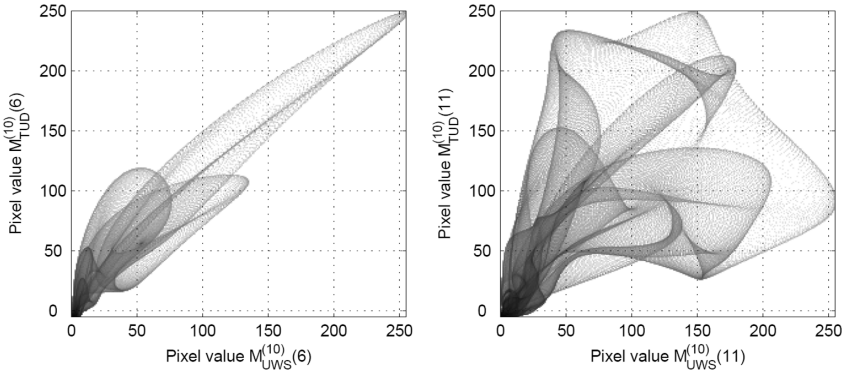


Figure 3.2.2: Example FDM for a presentation time of 10 s. Left to right: original, UWS, TUD, UN.

correlated FDM ($M^{(10)}_{UWS}(11)$ and $M^{(10)}_{TUD}(11)$) with $\rho_P = 0.637$), respectively. Naturally, the higher correlated FDM exhibit values much closer to the main diagonal. Nevertheless, there are also very distinct structures in the plots. The distinct structures may inherently result from the structures contained in the actual FDM (see Figure 3.2.2). However, the PLCC does not account for these structural differences between the FDM. Additionally, PLCC cannot distinguish whether differences amongst FDM are caused mainly from high magnitude pixels or low magnitude pixels. The area under the ROC curve accounts for these missing aspects of PLCC.



(a) (b)

Figure 3.3.1: Scatter-like plot of the conjoint pixel values between two FDM for (a) highly similar FDM ($M^{(10)}_{UWS}(6)$ and $M^{(10)}_{TUD}(6)$) and (b) highly dissimilar FDM ($M^{(10)}_{UWS}(11)$ and $M^{(10)}_{TUD}(11)$).

3.3.2 Area under the ROC curve (AUC)

To facilitate the use of the area under the ROC curve (AUC) [Fawcett 06] for measuring FDM similarity, one of the two FDM has to be thresholded into a binary map as

$$M^{(t)}_{bin,DB}(i) = \begin{cases} 1 & for\ M^{(t)}_{DB} > \tau \\ 0 & for\ M^{(t)}_{DB} \leq \tau \end{cases}$$

with $\tau \in [0 \ldots 254]$ and $DB \in \{UN, UWS, TUD\}$. For the computation of AUC, the original FDM values are linearly transformed from the range [0,1] to the range [0,255].

Note that ROC analysis is non-symmetrical. Depending on which FDM of the pair is used to create the binary map, the value of the resulting AUC can vary. We therefore compute the average over the two non-symmetrical AUC. Depending on the threshold chosen, different properties of the FDM can be analyzed. For a low threshold, the binary map covers a larger area than for a large threshold. Hence, for low threshold values we

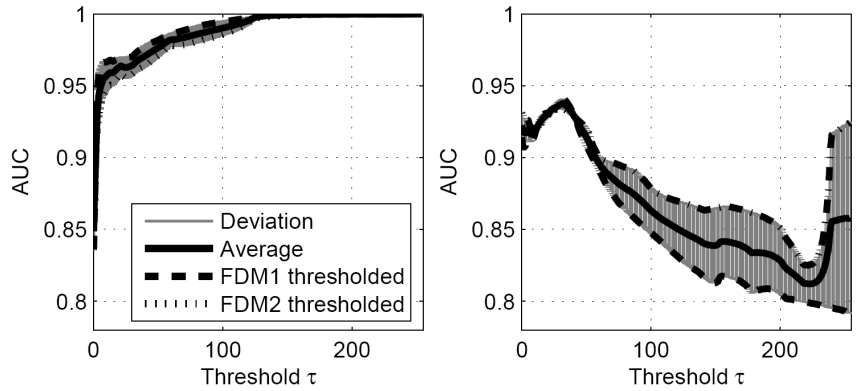


Figure 3.3.2: AUC for 255 thresholds τ between for (a) highly similar FDM ($M_{UWS}^{(10)}(6)$ and $M_{TUD}^{(10)}(6)$) and (b) highly dissimilar FDM ($M_{UWS}^{(10)}(11)$ and $M_{TUD}^{(10)}(11)$).

gain knowledge mainly about the coverage similarity between the FDM; for very large threshold values, we identify the similarity between the peaks of the FDM.

For illustration, Figure 3.3.2 presents both non-symmetrical AUC computations between two FDM, along with their mean for all 255 thresholds. For highly similar FDM (Figure 3.3.2(a)), the AUC rises fast towards the maximum level and the difference between the AUC is small. For highly dissimilar FDM (Fig. 3.3.2(b)), the AUC is low and in this case even decreases when the threshold increases. These lower AUC for large thresholds indicate that FDM $M_{UWS}^{(10)}(11)$ and $M_{TUD}^{(10)}(11)$ have different peaks, as can be visually observed from Figure 3.2.2.

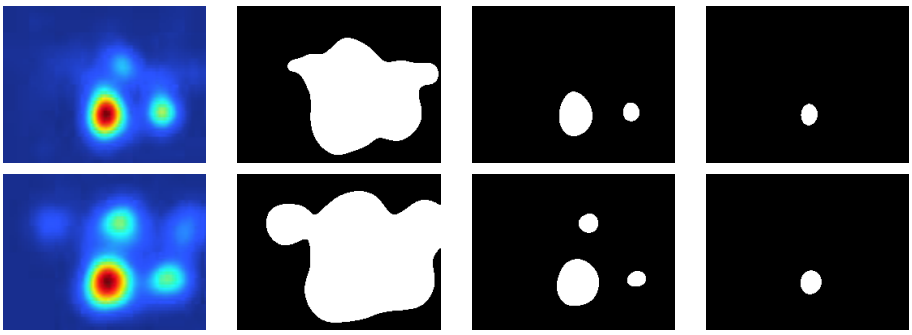


Figure 3.3.3: Binary maps for highly similar FDM ($M_{UWS}^{(10)}(6)$ and $M_{TUD}^{(10)}(6)$) after thresholding with $\tau = 10$, $\tau = 100$, and $\tau = 200$.

To capture different properties of FDM (e.g. coverage and peak), we consider in the following three different thresholds: a low threshold $\tau = 10$, a high threshold $\tau = 200$, and also an intermediate threshold $\tau = 100$ to account for lower order peaks. Figure

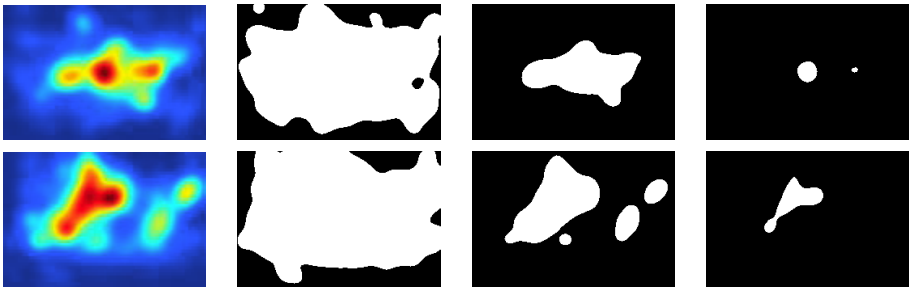


Figure 3.3.4: Binary maps for highly dissimilar FDM ($M^{(10)}_{UWS}(11)$ and $M^{(10)}_{TUD}(11)$) after thresholding with $\tau = 10$, $\tau = 100$, and $\tau = 200$.

3.3.3 and Figure 3.3.4 illustrate the binary maps resulting from the thresholding for the FDM presented in Figure 3.3.2. The similarity between the binary maps in Figure 3.3.3 reflects well the increase in AUC as presented in Figure 3.3.2(a). Similarly, the decrease in AUC in Figure 3.3.2(b) is also reflected in the visual inspection of the binary maps in Figure 3.3.4.

3.3.3 Monotonicity between PLCC and AUC

Although the purposes of performing the PLCC and AUC are different, both measures are expected to vary conjointly to some degree. To identify the degree to which the two measures interrelate to each other, we compute the Spearman rank order correlation coefficient (SRCC). The SRCC is computed over all images and presentation times t ($N = 29 \times 11 = 319$). The results are presented in Table 3.3. The results indicate that the threshold τ has a strong impact on the similarity between the ranks of PLCC and AUC. The SRCC for the thresholds $\tau = 100$ and $\tau = 200$ are higher than for the threshold $\tau = 10$. This result can be attributed to the fact that the PLCC is only high if also the large magnitudes in the FDM (the peaks) agree with each other.

	UN vs. UWS	UN vs. TUD	UWS vs. TUD
$\tau = 10$	0.474	0.409	0.317
$\tau = 100$	0.784	0.762	0.836
$\tau = 200$	0.64	0.762	0.691

Table 3.3: Spearman rank order correlation between PLCC and AUC.

3.4 Inter-laboratory comparison

The similarity of the FDM is in the following evaluated with regard to (1) all presentation times, (2) the impact of the central fixation point used or not in the experiments, and (3) the content dependency.

3.4.1 Inter-laboratory differences

The main objective of this section is to evaluate inter-laboratory differences between FDM. For this purpose, we illustrate the PLCC and AUC (see Figure 3.4.1 and Figure 3.4.2) as a function of presentation time t for the 3 inter-laboratory comparisons (e.g. UWS-TUD, UWS-UN, TUD-UN) and the comparisons to the random FDM (e.g. UWS-RND, TUD-RND, and UN-RND).

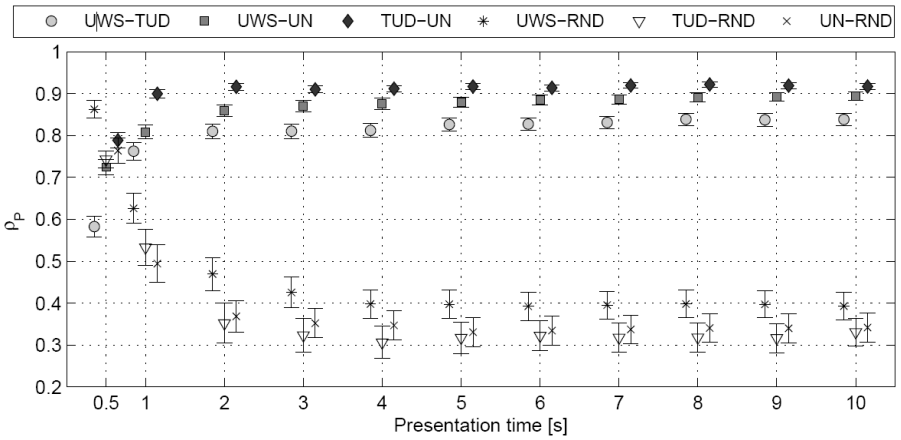


Figure 3.4.1: Mean PLCC and standard errors over all images for all t.

Each figure shows the means along with the standard errors over all the 29 images. Figure 3.4.1 illustrates that the progression of the mean PLCC with presentation time is similar between the three experiments. The increase quickly flattens out and the PLCC only marginally depends on the presentation time for $t \geq 2\,s$. For TUD-UN this observation already holds for $t \geq 1\,s$.

Despite the similar progressions of the PLCC, the overall magnitudes between the three comparisons differ to some degree. TUD-UN has the highest correlations, followed by UWS-UN and UWS-TUD. Hence, the FDM of experiments TUD and UN appear to be most similar, while their respective similarities to experiment UWS are to some degree lower. In addition to the lower mean PLCC, it can also be observed that the standard errors are larger for UWS-TUD and UWS-UN compared to TUD-UN, which indicates that there is a larger variance of the PLCC with respect to the image content. The significantly higher PLCC for $t \geq 1\,s$ between the experimental FDM, as compared to the random FDM, emphasize high similarity between the experimental FDM.

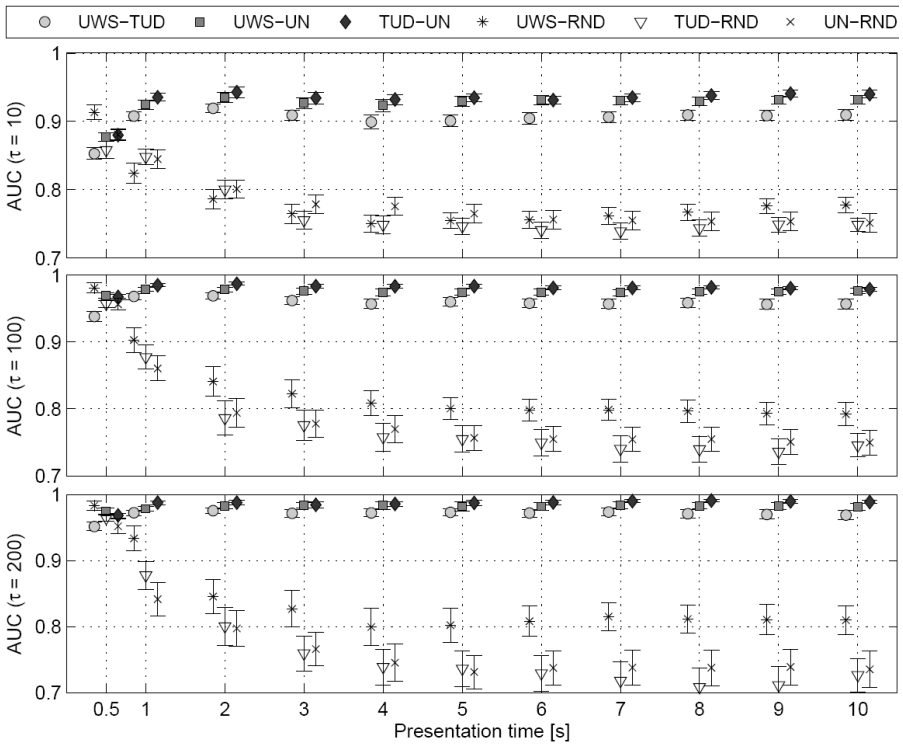


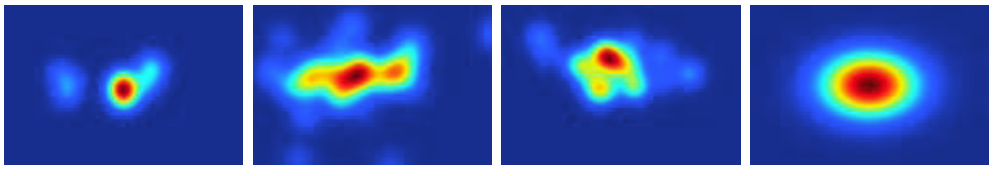
Figure 3.4.2: Mean AUC and standard errors over all images for all t and for $\tau = 10$ (top), $\tau = 100$ (middle), and $\tau = 200$ (bottom).

Similar observations as for the PLCC can be also found for the AUC (as presented in Figure 3.4.2). The AUC plots confirm the order of similarity between the three experiments and also the increasing similarity between experiments with an increase in presentation time. It is interesting to note the difference of the AUC values for the three different thresholds $\tau = 10$, $\tau = 100$ and $\tau = 200$: the AUC increases with the threshold. This phenomenon indicates that the similarity between experiments is generally higher for the strongly salient regions as compared to the other regions in the images. The significantly lower AUC between the experimental FDM and random FDM confirm the observations on the PLCC.

3.4.2 Impact of the central fixation point on the center bias

In Section 3.2.2, we discussed some possible factors that may impact the similarity of FDM from different experiments. One factor distinguishing UWS from TUD and UN is the central fixation point used in UWS. The use of a central fixation point can be expected to have an effect on the FDM due to a change of center bias during image viewing. Center bias describes the phenomenon that viewers of natural images tend to focus more on the central than the peripheral areas of the scene. Several factors are known to contribute to this effect, including, photographer bias, viewing strategy, orbital reserve, motor bias, and center of screen bias [Tseng 09]. It has been shown that center bias is predominant during free viewing tasks rather than visual search tasks [Tatler 07].

We analyse whether the central fixation point used in UWS affects the center bias, and consequently affects the similarity between the FDM. In Figure 3.4.3(a)-(c), we illustrate examples of FDM based on $t = 500\,ms$ presentation time for image number 18 ('parrots'). The central fixation point used in UWS resulted in a more centered distribution of the fixations. Despite the missing of a central fixation point, the fixations in UN and TUD also experienced a tendency towards the center of the image. However, the fixation distributions of UN and TUD are with a wider spread and an off-center shift. Similar observations hold for the whole image set.



(a) (b) (c) (d)

Figure 3.4.3: Comparison of FDM for $t = 500\,ms$: (a)$M_{UWS}^{(0.5)}(18)$, (b)$M_{UN}^{(0.5)}(18)$, (c)$M_{TUD}^{(0.5)}(18)$, and (d) center-bias map.

For each of the 29 images, we created individually a center bias map based on an anisotropic Gaussian kernel according to the procedures described in [Le Meur 06]. An

example center bias map is shown in Figure 3.4.3(d) for image number 18. We then compute the PLCC and AUC (with $\tau = 200$) between these center bias maps and the FDM. The result is plotted in Figure 3.4.4. The PLCC and AUC are consistently higher for UWS than for TUD and UN for almost all presentation times. Only PLCC for $t = 500\,ms$ contradicts these results, which may be attributed to the more widely spread fixations in TUD and UN that might correlate better with the widely spread center bias maps.

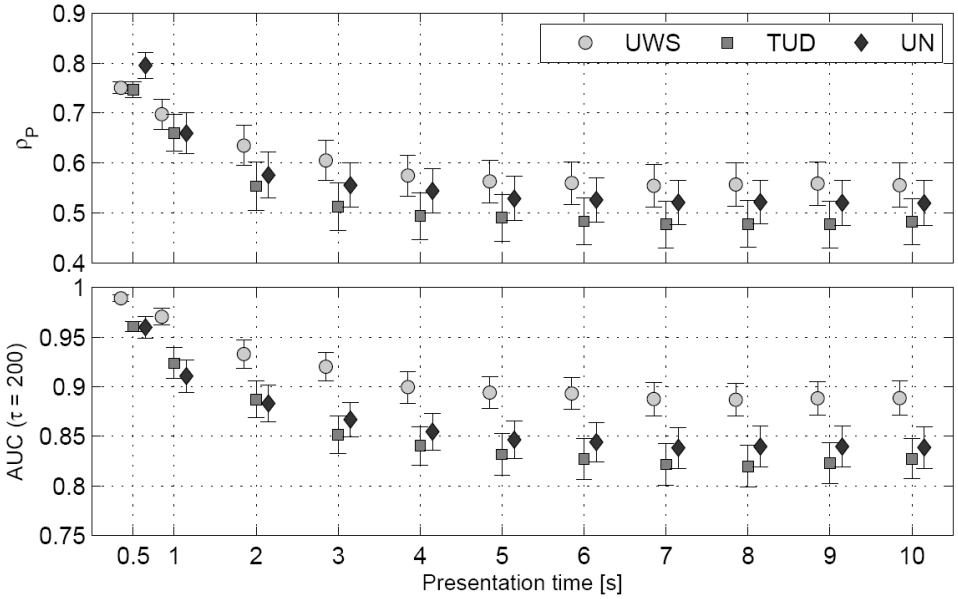


Figure 3.4.4: PLCC (top) and AUC for $\tau = 200$ (bottom) between the FDM and the respective center bias map for all presentation times t.

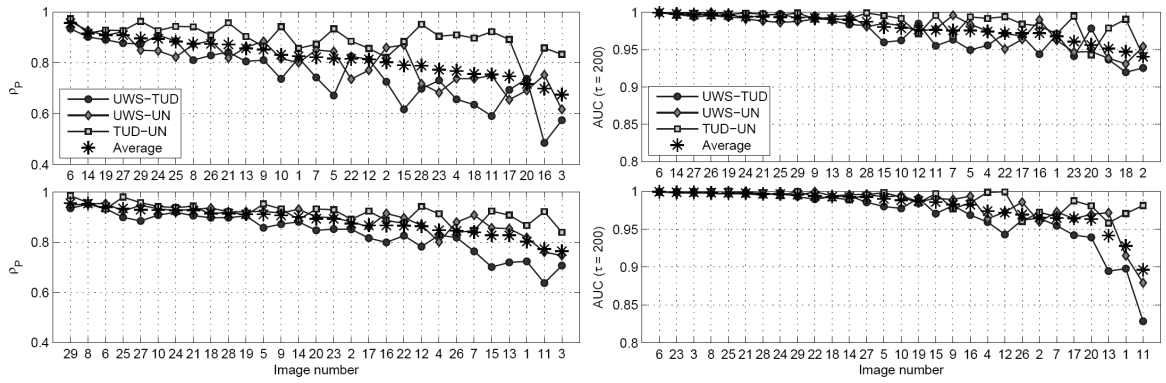
We performed paired t-tests at 95% confidence on the metric values presented in Figure 3.4.4 to verify whether the higher UWS values are statistically significant. The t-tests were performed for PLCC and AUC ($\tau = 200$) and for all presentation times. The results (see Figure 3.4) reveal if there exists a significant difference between the metric values. The results show that the AUC between UWS and the center bias maps (UWS-CB) are significantly higher than for the other comparisons (TUD-CB,UN-CB). On the other hand, the comparisons UN-CB and TUD-CB are not statistically different for AUC of all presentation times. The t-test between PLCC of (UWS-CB) and (TUD-CB) confirms these results for $t \geq 2s$. The PLCC values of (UWS-CB) show clearly a trend of higher values compared to (UN-CB) (see Figure 3.4.4), however, most differences are found not to be statistically significant. The PLCC of (TUD-CB) are significantly lower than for (UN-CB) for $t \geq 3s$ (although most are not far from insignificance with $p \in \{0.012...0.035\}$). Overall, these results reveal that the FDM of UWS indeed experience a stronger center bias compared to TUD and UN. We argue that the central fixation point in UWS is likely to be the cause for this difference.

	Paired t-test result (PLCC/AUC) for each presentation time t										
	0.5	1	2	3	4	5	6	7	8	9	10
UWS-CB, UN-CB	1/1	0/1	1/1	1/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
UWS-CB, TUD-CB	0/1	0/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1
UN-CB, TUD-CB	0/0	0/0	0/0	1/0	1/0	1/0	1/0	1/0	1/0	1/0	1/0

Table 3.4: Results of paired t-test at 95% confidence. '1' represents statistically significant different, whereas '0' represents no significant difference. In each grid, the left number indictes the paired t-test result for PLCC; the right number indicates the paired t-test result for AUC ($\tau = 200$).

3.4.3 Content dependency

The standard errors in Figure 3.4.1 and Figure 3.4.2 show that the FDM similarity is to some degree content dependent. We therefore analyse here the similarity between the FDM in relation to the content of the images. Given that the PLCC and AUC are very similar for $t \geq 2s$, we consider two presentation times: $t = 1s$ and $t = 10s$. These two presentation times allow us to compare the influence of image content on both the early fixations ($t = 1s$) and on a more exhaustive viewing of the images ($t = 10s$). In Figure 3.4.5, the PLCC and AUC for all 29 images are presented. The PLCC and AUC are computed individually for the three comparisons amongst experiments (i.e. UWS-TUD, UWS-UN and TUD-UN). Moreover, the average value over the three comparisons is also computed. Note that all values presented in Figure 3.4.5 are sorted with respect to the decreasing average measures.



(a) (b)

Figure 3.4.5: Impact of the image content measured using (a) PLCC and (b) AUC ($\tau = 200$) for $t = 1s$ (top) and $t = 10s$ (bottom).

It is found that the similarity amongst FDM appears to strongly depend on the image content, both for $t = 1s$ and $t = 10s$. For high average PLCC and AUC, the values from

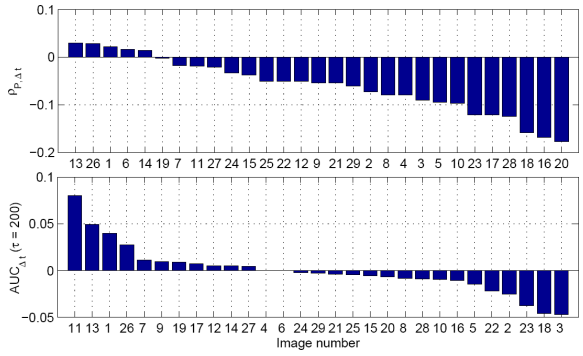


Figure 3.4.6: Differences between $t = 1s$ and $t = 10s$ for PLCC (top), $\rho_{P,\Delta t}$, and AUC with $\tau = 200$ (bottom), $AUC_{\Delta t}$.

the individual experimental comparisons are located closely together, whereas for low average PLCC and AUC the deviation of the individual values is considerably higher.

One can further observe that the PLCC and AUC of the same images can be different between the two presentation times. This essentially means that the content dependency of the similarity between FDM is a function of time. To better illustrate the differences between the two presentation times, we present in Figure 3.4.6 two bar plots of the PLCC and AUC difference between $t = 1s$ and $t = 10s$ denoted as $\rho_{P,\Delta t}$ and $AUC_{\Delta t}$, respectively. Together with Figure 3.4.5 as well as visual inspection of the FDM at all presentation times, these results allow for a more detailed discussion as follows.

For both PLCC and AUC, many of the images exhibit rather small differences $\rho_{P,\Delta t}$ and $AUC_{\Delta t}$. For instance, image number 6 ('cemetry') is rated very high for both $t = 1s$ and $t = 10s$. This image contains two plaques with written text which attracted the attention of the observers upon presentation of the image and kept the attention throughout the image presentation. Many other images have large differences between presentation times. Image number 3 ('buildings'), for instance, exhibits the largest $AUC_{\Delta t}$ difference in the set due to a low AUC at $t = 1s$ and a high AUC at $t = 10s$. Like image number 6, this image also contains written text. However, due to the high complexity of the remainder of the image, the text is not as dominant and the observers needed more time to detect it. Similarly, image number 18 ('parrots') has considerably higher PLCC and AUC for $t = 10s$ as compared to $t = 1s$. This image contains two distinct salient regions (the parrot heads) which attract, respectively, a comparable amount of attention with increased presentation time.

3.5 Impact on applications

One of the main purposes of conducting eye-tracking experiments and creating experimental FDM is to provide reliable ground truth to the community. The similarity measures, PLCC and AUC, indicate high similarity between the FDM from different

experiments. However, they do not provide direct insight into the reliability of these FDM as a ground truth for image processing applications. In this section, we therefore identify the sensitivity of three applications to the FDM used: visual saliency modelling, image quality assessment and image retargeting.

3.5.1 Impact on performance of visual saliency models

FDM obtained from eye-tracking experiments are typically used for the training and validation of visual saliency models. We analyse here to what degree the validation of saliency models depends on the ground truth, the FDM. We consider in the following the well known saliency model by Itti et al. [Itti 98] as well as the models by Rajashekar et al. [Rajashekar 08], Bruce et al. [Bruce 09], Achanta et al. [Achanta 09], and Hou et al. [Hou 07]. We compute the saliency maps for all images using these models and compute the similarity between them and the FDM based on presentation times $t = 1\,s$ and $t = 10\,s$. The results for PLCC and AUC ($\tau = 200$) are presented in Table 3.5. In addition, the standard deviations over the PLCC and AUC are given over the three databases, σ_{DB}, and the five saliency models, σ_{SAL}.

The results show that both similarity measures, PLCC and AUC, differ considerably more between the visual saliency models than between the FDM. This observation holds for both presentation times $t = 1\,s$ and $t = 10\,s$. Note that all saliency models perform better on the FDM with $t = 10\,s$, even though these models aim to predict salient regions that are believed to be driven by rapid bottom-up mechanisms. For the PLCC, this higher performance might be influenced to some degree by the larger number (and thus a wider spread) of fixations for $t = 10\,s$ compared to $t = 1\,s$. The results also show the consistently higher performance of all saliency models on the experimental FDM compared to the random FDM. It further illustrates that the models predict saliency with an accuracy above chance.

3.5.2 Impact on performance of image quality assessment

Saliency maps and FDM are often integrated into image quality metrics with the aim to improve quality prediction performance [Hantao 11]. We analyze to what degree the improvement of image quality metrics varies with the FDM used. For practical reasons, the objective metrics used in our evaluation are limited to three full-reference metrics that are so far widely accepted in the image quality community: PSNR (peak signal-to-noise ratio [Wang 06]), SSIM (structural similarity index [Wang 04]) and VIF (visual information fidelity [Sheikh 06]). A combination strategy which has been proven to be efficient is used in our work. Following the procedure in [Liu 11], we integrate the saliency map into the objective quality model by doing local and multiplicative weighting of the respective distortion map. An example is illustrated in Figure 3.5.1. The addition of saliency to PSNR, SSIM and VIF results in three attention-based metrics, which are referred to as WPSNR, WSSIM, and WVIF. They can be defined as follows:

t	Database	Visual attention models					σ_{SAL}
		Itti	Rajashekar	Bruce	Achanta	Hou	
1s	UN	0.099	0.372	0.272	0.254	0.32	0.103
	UWS	0.096	0.288	0.218	0.202	0.241	0.071
	TUD	0.097	0.348	0.244	0.242	0.300	0.094
	σ_{DB}	0.001	0.044	0.027	0.027	0.041	—
	RND	0.040	0.282	0.164	0.120	0.152	0.087
10s	UN	0.150	0.449	0.376	0.335	0.421	0.118
	UWS	0.147	0.435	0.371	0.312	0.384	0.111
	TUD	0.152	0.448	0.369	0.333	0.415	0.115
	σ_{DB}	0.003	0.008	0.004	0.013	0.020	—
	RND	0.072	0.297	0.208	0.135	0.171	0.084

(a)

t	Database	Visual attention models					σ_{SAL}
		Itti	Rajashekar	Bruce	Achanta	Hou	
1s	UN	0.648	0.801	0.749	0.681	0.758	0.061
	UWS	0.624	0.733	0.683	0.651	0.676	0.040
	TUD	0.620	0.786	0.713	0.693	0.748	0.062
	σ_{DB}	0.015	0.036	0.033	0.022	0.044	—
	RND	0.621	0.717	0.633	0.603	0.650	0.044
10s	UN	0.658	0.802	0.773	0.690	0.760	0.060
	UWS	0.660	0.797	0.758	0.687	0.737	0.055
	TUD	0.671	0.803	0.772	0.692	0.771	0.057
	σ_{DB}	0.007	0.003	0.008	0.002	0.017	—
	RND	0.616	0.682	0.624	0.554	0.606	0.046

(b)

Table 3.5: (a) PLCC and (b) AUC ($\tau = 200$) between the predicted saliency maps and FDM.

$$WMetric = \frac{\sum_{x=1}^{M}\sum_{y=1}^{N}[D(x,y)\cdot S_i(x,y)]}{\sum_{x=1}^{M}\sum_{y=1}^{N} S_i(x,y)}$$

where D represents the distortion map calculated by the given metric, S indicates the saliency map used, and $WMetric$ denotes the resulting attention-based metric (e.g. WPSNR, WSSIM and WVIF).

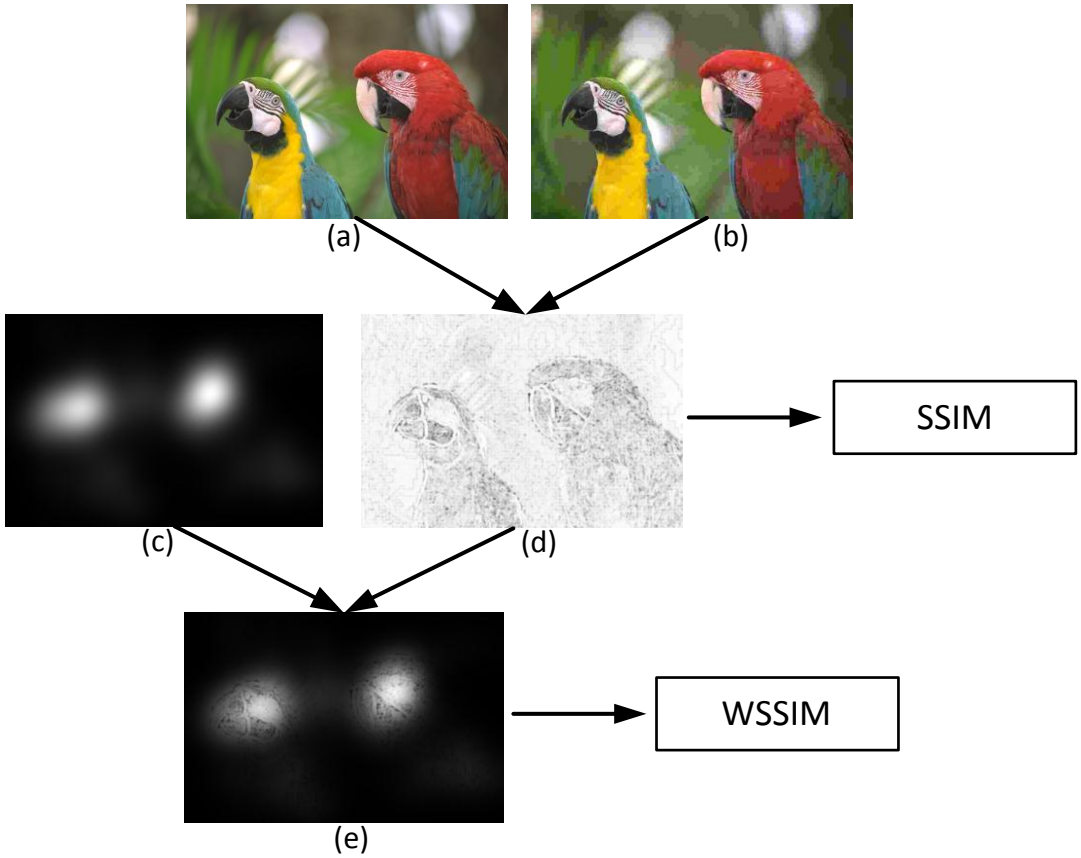


Figure 3.5.1: Illustration of an objective metric based on saliency. (a) original image. (b) The JPEG compressed image. (c) The corresponding fixation density map showing NSS, the lower the intensity, the lower the attention. (d) Distortion map of SSIM calculated for (a) and (b). (e) Combination of distortion map and NSS, the lower the intensity, the larger the distortion is.

To analyze impact of the inter-database difference, we computed PSNR, SSIM, and VIF on all distorted images of the LIVE database before and after incorporation of the FDM (i.e. WPSNR, WSSIM and WVIF) collected from the three different experiments.

We integrate the FDM based on a presentation time of $t = 10s$. As the images we used in our eye tracking experiments are taken from the LIVE image quality database, we have a large set of distorted images and their respective mean opinion scores (MOS) available for the design and validation of the quality models. The metrics' predictions were compared to the respective MOS by computing the PLCC. The analysis is conducted independently for the different distortion classes contained in the LIVE database. The performance gains of different quality metrics through incorporation of the FDM of different databases are presented in Table 3.6, Table 3.7 and Table 3.8.

		JPEG 1	JPEG 2	J2K 1	J2K 2	Gaussian blur	White noise	Fast fading	Average
PSNR	UWS	0.004	0.028	0.002	0.037	0.006	0	0.009	0.012
	TUD	0.008	0.031	0.006	0.037	0.020	0	0.016	0.017
	UN	0.006	0.029	0.003	0.036	0.018	0	0.015	0.015
	RND	0	-0.006	-0.001	0.005	-0.02	0.014	0.001	-0.008

Table 3.6: PSNR prediction performance gain based on PLCC.

		JPEG 1	JPEG 2	J2K 1	J2K 2	Gaussian blur	White noise	Fast fading	Average
SSIM	UWS	0.017	0.041	0.022	0.041	0.075	0.008	0.029	0.033
	TUD	0.019	0.039	0.019	0.038	0.070	0.008	0.023	0.031
	UN	0.014	0.036	0.019	0.037	0.070	0.009	0.025	0.030
	RND	0.002	-0.004	0.002	0	0.064	-0.002	0.023	0.012

Table 3.7: SSIM prediction performance gain based on SSIM.

		JPEG 1	JPEG 2	J2K 1	J2K 2	Gaussian blur	White noise	Fast fading	Average
VIF	UWS	0.022	0.008	0	0.007	0.017	0.010	0.012	0.011
	TUD	0.022	0.008	0	0.009	0.017	0.012	0.009	0.011
	UN	0.026	0.008	0.004	0.009	0.021	0.010	0.008	0.012
	RND	0.018	0.007	0	0.007	0.013	0.008	0.003	0.008

Table 3.8: VIF prediction performance gain based on VIF.

The results show that for all distortion classes and the related average, the improvements are very similar between the three experiments. The improvement, however, differs considerably between the quality prediction models as well as between the different distortion types. Hence, the results show that the consistency between FDM is

better than the consistency between quality prediction models. The consistently larger gain by using the experimental FDM compared to the random FDM further shows that the incorporation of FDM into quality prediction models is indeed beneficial for LIVE database and the three quality metrics selected.

3.5.3 Impact on performance of saliency-based image retargeting

Image retargeting algorithms [Avidan 07] resize images by cutting out vertical seams of lowest energy, thus preserving the most important regions in the images. Saliency-based image retargeting algorithms allow for an additional importance weighting based on the visual saliency in the scene. We used the FDM based on $t = 1\,s$ and $t = 10\,s$ in the saliency-based image retargeting algorithm by Wang et al. [Wang 11a] to investigate the similarity of the resulting retargeted images. Examples are presented in Figure 3.5.2 for the images 27, 29, and 13 of the LIVE database.

The retargeting fails for both presentation times when using the randomly substituted FDM (i.e. RND), which is particularly true for images 27 and 29. However, the most relevant regions are well preserved when using the experimental FDM, regardless the FDM are from UWS, TUD or UN. The outcomes are very similar between the databases. The similarity is particularly high between TUD and UN and is somewhat lower for UWS. This difference confirms our earlier results on PLCC and AUC in Section 3.4.

3.6 Discussion

3.6.1 Inter-laboratory differences

In Section 3.4, we discussed the differences between the three experiments and how they can be expected to have an impact on the similarity between the FDM. Given the multitude of varying factors due to the independently conducted experiments, we could only speculate here as to what degree each of the factors influences the inter-laboratory differences. We can, however, summarize the quantitative results from the PLCC and AUC metrics as well as from the three applications to observe some overall trends.

Between the three databases, both the PLCC and AUC show similar progressions with presentation time (see Figure 3.4.1 and Figure 3.4.2). The absolute values, however, are not exactly the same for the individual comparisons between the databases: TUD-UN is most similar, followed by UWS-UN and UWS-TUD. This trend is transferred to some degree to the image processing applications, where UWS typically differs somewhat more from the other databases. To see whether there are significant differences between the individual comparisons, we performed paired t-tests for all individual comparisons and all presentation times on the data presented in Figure 3.4.1 and Figure 3.4.2. The markers in Figure 3.6.1 reveal if there are significant differences between the comparisons at 95% confidence. For PLCC, all comparisons are statistically different, while for AUC only the

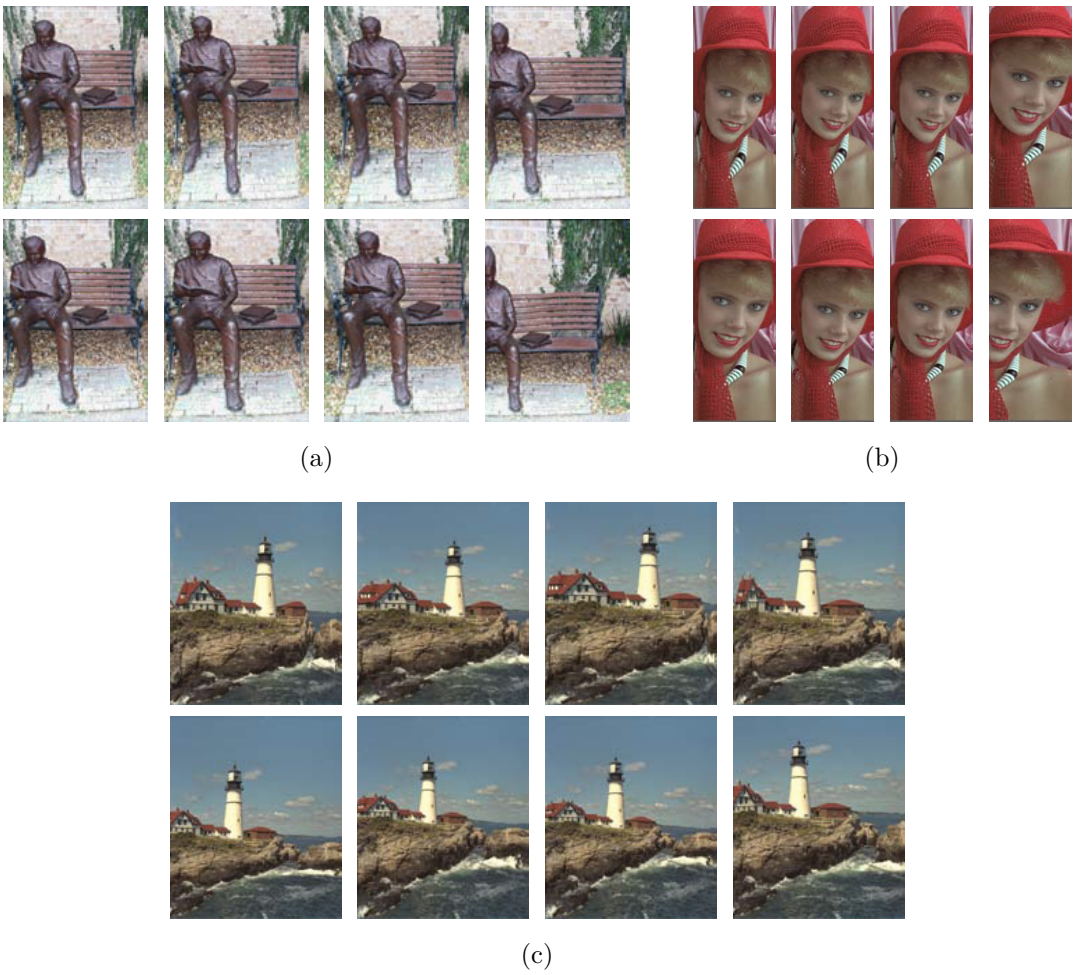


Figure 3.5.2: Retargeting result based on FDM with $t = 1\,s$ (the top row) and $t = 10\,s$ (the bottom row) for (a) Image 27, (b) Image 29, and (c) Image 13. From left to right: UWS, TUD, UN, RND.

comparisons UWS-TUD and UWS-UN are statistically the same in most cases. These results show indeed that inter-laboratory comparisons can vary significantly depending on the laboratories that are involved. Despite the significant differences regarding similarity metrics, the experimental FDM still have a similarly positive impact on the image processing applications, as compared to using randomly chosen FDM.

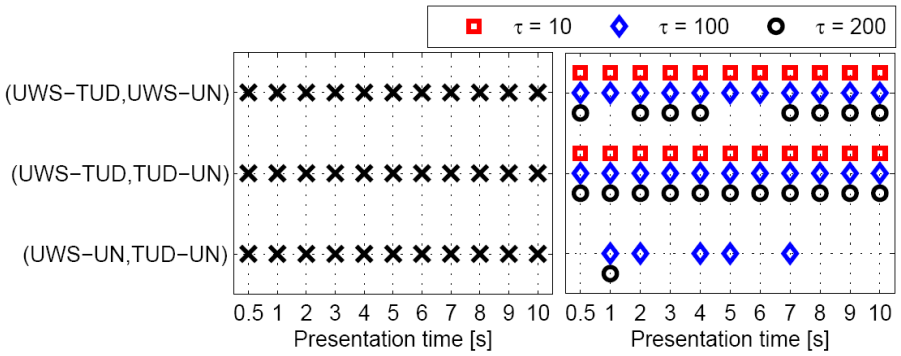


Figure 3.6.1: Paired t-test at 95% confidence for PLCC (left) and AUC (right). Markers indicate statistically significant difference.

3.6.2 Intra- versus inter-experiment differences

It could be argued that the differences between experiments are due to intrinsic variations amongst the observer groups. We therefore take a closer look at the variations within the experiments (intra-experiment), compared to the variations between the experiments (inter-experiment). We adopt the performance efficiency method [Stankiewicz 11] by repeatedly splitting the observer group within an experiment into two sub-groups and computing the PLCC between the FDM created from these groups. These PLCC are considered as the upper theoretical limit of the performance of saliency models. They serve also as an intrinsic ground truth of the variations amongst observers within an experiment. To facilitate a fair comparison, we adapt the method in [Stankiewicz 11] by selecting the same size of the sub-groups for the comparisons within and between the experiments. Based on these sub-groups, we create FDM for a presentation time of $t = 10s$. For the sub-group selection, we are bound by the lowest number of 15 observers in experiment UWS. We therefore randomly select two groups of 7 observers within each of the experiments and compute the intra-experiment PLCC between the related FDM. Similarly, we select randomly 7 observers from each experiment and compute the inter-experiment PLCC. To obtain a robust estimate we repeat this process 100 times for the intra- and inter-experiment comparisons and compute the average PLCC.

All intra- and inter-experiment PLCC are presented in Figure 3.6.2. The intra-experiment correlation for UN is approximately 5% higher than for UWS and TUD. One could speculate that the larger foveal coverage in relation to the image size in UN

(see Section 3.4) may enhance observers to grasp the gist of the scene. Thus, the number of possible target objects is lower and agreement between observers is higher. The higher accuracy of the eye-tracker used in UN could also have an impact on these results. Finally, the larger Gaussian kernel sizes relative to the image size inherently increases the PLCC to some degree.

Regarding the inter-experiment correlations, the PLCC are considerably higher for UWS-UN and TUD-UN than for UWS-TUD. The lower intrinsic differences of observers within UN may be one reason why this experiment generally correlates higher with the other experiments.

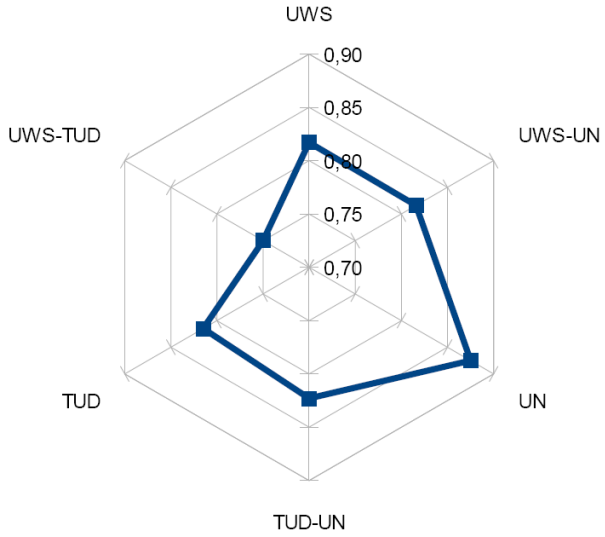


Figure 3.6.2: Spider chart of the intra-experiment (UWS, TUD, UN) and interexperiment (UWS-TUD, UWS-UN, TUD-UN) correlations between FDM based on 7 observers (values averaged over 100 random samples).

3.7 Conclusions and perspectives

In this chapter, we analyzed FDM similarity between three independent eye-tracking experiments, using two different similarity measures: PLCC and AUC. We showed that these measures capture different properties while being coherent in predicting the similarity of the FDM. Only for short presentation times ($t \leq 1s$), PLCC was found to deviate from AUC. Despite various differences between the experiments, the FDM were found to be very similar. The degree of similarity, however, was dependent on the individual experimental comparisons: UWS is more different to TUD and UN. The FDM similarity was further revealed to be highly dependent on the image content. Images that contain a distinct salient region experience a higher FDM similarity as compared to images with multiple or no salient regions. The reliability of the FDM as a ground

truth has also been validated on three image processing applications: visual saliency modeling, image quality assessment and image retargeting. On all the applications it was shown that the difference between the experimental FDM on the outcomes was low. These findings suggest that FDM from independent eye-tracking experiments can indeed be considered as reliable ground truths for image processing applications.

Given the independence of the experiments, we could not quantify the degree to which the various factors are affecting the FDM. It is therefore instrumental to extend this work by conducting experiments conjointly with careful variation of certain factors to evaluate their impact on the FDM. It further needs to be verified whether a larger number of participants would result in even more stable FDM and thus in higher similarity between the experiments. Thresholds need to be determined that specify the minimum number of participants in order to achieve FDM for given similarity constraints. Finally, the comparisons presented in this study hold for eye-tracking experiments under task-free condition. Different results could be expected under a variety of viewing tasks, for instance, visual search tasks. These issues are out of the scope of this study and are subject for future work.

The degree to which fixation distribution can represent the higher level of attention remains an open question. In our study, the results exhibit a strong relationship between the FDM and the content of the image, as well as the time-dependency of this relationship. However, during free viewing, the fixation distribution is traditionally considered to be linked with bottom-up stimuli, such as color, intensity. On the other hand, it is the top-down mechanisms which are linked with the content of the image, e.g. gist of the scene and object recognition. A study of the quantification of this relationship between fixation distribution and top-down mechanisms is thus of particular importance. In the next chapter, we will introduce such a study.

Key points

Context

- ❏ Fixation density map (FDM) obtained by eye-tracking experiment is believed to provide reliable ground truth of overt visual attention. Since conducting eye-tracking experiment is expensive, several FDM database for image or video have been made publicly available to the society.
- ❏ Most of eye-tracking experiments are conducted independently in different laboratories; the procedures of the experiments can be also different. Nevertheless, the outcomes of these experiments, the FDM, are considered to be solid ground truths.
- ❏ The knowledge about the degree to which these "ground truth" differs between experiment, and its impact on the applications of FDM, is of great importance. However, to the best of our knowledge, there has not been any comprehensive study on the comparison of eye tracking data amongst different laboratories.

Contributions

- ❏ We perform a thorough comparison of FDM that were created from three independently conducted eye-tracking experiments. We particularly focus on the effect of presentation time and image content.
- ❏ In addition to the analyses on FDM themselves, we also evaluate the impact of the FDM differences on three applications: computation modeling of visual saliency, image quality assessment and image retargeting.
- ❏ Our results show that the FDM from different experiments are very similar. The impact of the difference on the three applications studied (i.e. saliency modeling, image quality assessment and image retargeting) is low for the LIVE database.

.

Chapter 4

Linking visual salience and visual importance

In the previous chapter, we examined the reliability of fixation density maps (FDM), which are usually used as ground truth since they are generally believed to represent the salience of different areas in a scene. However, it is worth to note that FDM is not the only type of ground truth that has been used in attention-based applications. One might have noticed that, when visual attention is taken into account by the signal-processing community, another term "importance" (or "interest") have traditionally been considered synonymous with salience. In the literature, both saliency map and importance map (i.e. regions-of-interest) have ever been used as ground truth representing the most visually "relevant" parts of the scene.

From the vision science point of view, what differentiates visual salience and visual importance might be their intrinsic attributes that are related to bottom-up and top-down mechanisms. Nevertheless, the quantitative relationship between these two types of ground truth remains as an open question in signal-processing community. Therefore, we present in this chapter a psychophysical study[1] which concerns the quantification of the relationship between visual salience and visual importance.

This chapter is organized as follows: Section 4.2 describes the experimental methods used to collect the saliency maps and importance maps. Section 4.3 presents the results and analyses of the experiments. A discussion of our findings is provided in Section 4.4. General conclusions are presented in Section 4.5.

[1]This study is performed through an international collaboration with Computational Perception and Image Quality Lab, Oklahoma State University, USA.

4.1 Introduction

Visual salience [Koch 85, Itti 98] and visual importance [Osberger 98, Maeder 95, Etz 00, Kadiyala 08] come from the two different mechanisms of visual attention, the bottom-up mechanism and the top-down mechanism, respectively. Both visual salience and visual importance can provide important insights into how biological visual systems address the image-analysis problem. Both of them are also believed to denote the most visually "relevant" parts of the scene. However, despite the differences in the way (bottom-up) visual salience and (top-down) visual importance are determined in terms of human visual processing, both salience and importance have traditionally been considered synonymous in the signal-processing community.

In this chapter, a study measuring the similarities and differences between visual salience and visual importance is presented. We present the results of two psychophysical experiments and the associated computational analyses designed to quantify the relationship (and its evolution over time) between visual salience and visual importance:

- A psychophysical experiment was performed to obtain visual importance maps for a large database of images. A visual importance map is an object-level map that specifies the visual importance of each object in an image relative to the other objects in the image (including what would normally be considered as the background). The object(s) that receive the greatest visual importance are traditionally considered as the image's main subject(s). By using images from the Berkeley Image Segmentation Dataset, we collected importance ratings for each object in the 300 database images. Such importance ratings are generally believed to result from top-down visual processing since the decisions used to rate each object typically involve scene interpretation, object recognition, and often consideration of artistic intent.

- In a second experiment, visual gaze patterns were measured for 80 of the images from the same Berkeley Image Segmentation Dataset. Using an eye-tracker, visual gaze locations were recorded under task-free viewing. Whereas importance maps are driven primarily by top-down processing, visual gaze patterns are generally believed to be driven by bottom-up, signal-based attributes, at least for early gaze locations. Bottom-up saliency [Koch 85] is one particular signal-based attribute which has been shown to correlate well with early gaze locations. An image region is considered visually salient if it "stands out" from its background in terms of one or more attributes (e.g., contrast, color, orientation). When visual gaze patterns are measured in task-free viewing, one can consider the locations to denote the salient regions in the image. Thus, from the gaze patterns, one can construct an experimental saliency map.

The use of the same images in both experiments allows us to perform a computational analysis of the saliency maps and importance maps to quantify their similarities and dif-

ferences. A similar analysis was earlier performed by Engelke et al. in which visual fixation patterns were compared to human-selected regions of interest (ROI) [Engelke 09b]. Engelke et al. concluded that there indeed exists a relationship between visual fixation patterns and ROI, with early fixations demonstrating a stronger relationship than later fixations. Here, we perform a related study using importance maps rather than human-selected ROIs, with a particular emphasis on quantifying any potential relationships as a function of time.

4.2 Methods

Two psychophysical experiments were performed to obtain saliency maps and importance maps. In Experiment I, subjective ratings of importance were recorded to obtain importance maps. In Experiment II, visual gaze patterns were recorded to obtain saliency maps.

4.2.1 Experiment I: Visual Importance

In Experiment I, subjective ratings of importance were obtained for each object in each of 300 images to obtain importance maps. The methods were as follows.

Apparatus

Stimuli were displayed on a ViewSonic VA912B 19-inch LCD monitor (1280×1024 at 60 Hz). The display yielded minimum and maximum luminance of respectively, 2.7 and 207 cd/m^2. Stimuli were viewed binocularly through natural pupils in a darkened room at a distance of approximately 1575 pixels.

Stimuli

Stimuli used in Experiment I were obtained from the Berkeley Segmentation Dataset and Benchmark image database. This database was chosen because its images are accompanied by human-segmented versions (averaged over at least five subjects). We hand-segmented all 300 images in the database into 1143 objects by using the database's hand-segmented results as a reference (the database provides only edge-map segmentations rather than object-map segmentations). The images used were 321×481 or 481×321 pixels with 24-bit RGB values.

Procedures

For each of the 1143 objects, subjects were instructed to rate the perceived importance relative to the other objects within the image. The ratings were performed using an

integer scale of 0 to 10 in which 10 corresponded to greatest importance and 0 corresponded to least importance. The time-course of each session was not limited; however the majority of subjects completed the experiment in less than 120 minutes.

Raw scores for each subject were converted to z-scores. The per-subject z-scores were then averaged across all subjects, and then the average z-scores were rescaled to span the range $[0, 1]$ for each image. From these results, a per-image *importance map* was obtained by assigning each object's average importance score to all pixels in that object. Thus, in each importance map, brighter regions denote objects of greater visual importance.

Subjects

Ten adult subjects participated in the experiment. Three of the subjects were familiar with the purpose of the experiment; the other subjects were naive. Subjects ranged in age from 21 to 34 years. All had either normal or corrected-to-normal visual acuity.

4.2.2 Experiment II: Visual Salience

In Experiment II, an eye-tracker was employed to measure visual gaze points and thereby to compute visual saliency maps using 80 of the images used in Experiment I. The methods were as follows.

Apparatus

Stimuli were displayed on a Dell 1704FPT LCD 17-inch monitor (1280×1024 at 60 Hz). The display yielded minimum and maximum luminance of respectively, 0.3 and 180 cd/m^2. Stimuli were viewed binocularly through natural pupils in a darkened room at a distance of approximately 1575 pixels through natural pupils. Eye-tracking was performed by using the Video Eyetracker Toolbox from Cambridge Research Systems. This system tracks via acquisition/processing of pupil and dual first Purkinje images. The system has an accuracy of 0.25° and a sampling rate of 50 Hz.

Stimuli

The stimuli used in Experiment II consisted of 80 of the 300 images used in the Experiment I. The 80 images were selected to contain only the landscape orientation (481×321), and to provide a uniform sampling in terms of the number of objects per image and the levels of importance per image (see Table 4.1).

Procedures

A task-free viewing paradigm was employed in which subjects were instructed simply to look at the images given no specific task. Each of the 80 stimuli was presented for 15 seconds. The order of the presentation was randomized for each observer. Calibration

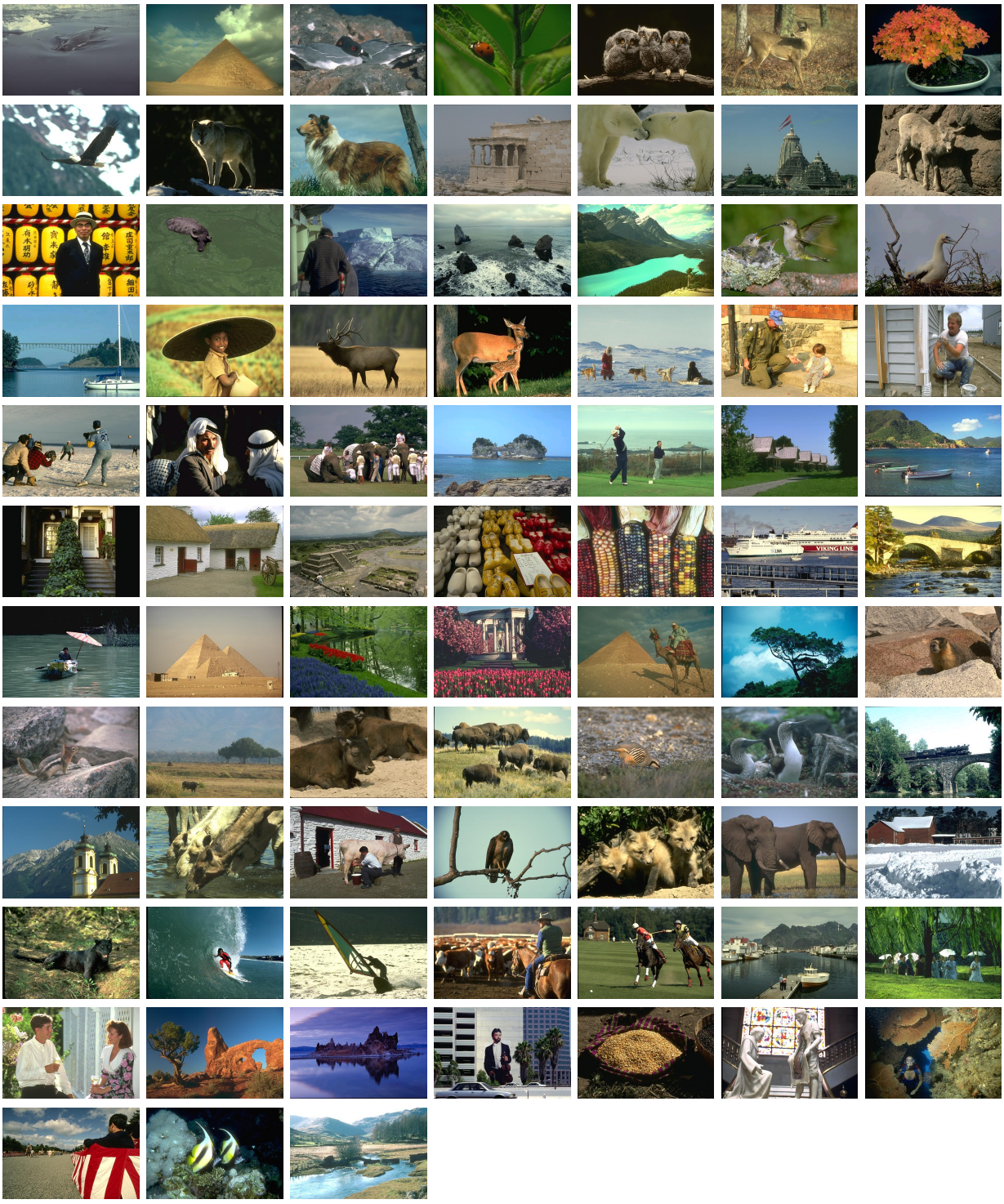


Table 4.1: The eighty images used in both our experiments to collect visual importance map and visual saliency map.

of the eye-tracker was performed periodically throughout the approximately 30-minute experimental session.

From the gaze position samples, we constructed a per-image *saliency map* by placing a two-dimensional Gaussian at each gaze sample point. The standard deviation of the Gaussian was determined based to the size of fovea (i.e. 1 degree of visual angle). We then normalized all saliency maps to span the range $[0, 1]$.

Subjects

Eighteen adult subjects participated in the experiment. All subjects were paid participants and were naive to the purpose of the experiments. Subjects ranged in age from 19 to 45 years. All had either normal or corrected-to-normal visual acuity.

4.3 Results and analysis

4.3.1 Qualitative Observations of Importance Maps and Saliency Maps

A qualitative comparison of the saliency maps and importance maps reveals some distinct similarities and differences between the two. Figure 4.3.1 depicts some representative examples.

The importance maps suggest that object category plays a bigger role than most other factors in determining subjective importance. In general, we found that subjects tended to rate objects containing human faces and/or animals to be of greatest importance. Background objects such as sky and grass were generally rated to be of least importance. Occlusion (i.e., whether an object is in the foreground vs. the background) also seems to be an important factor for perceived importance.

The saliency maps generally suggest that regions which possess a distinguished shape, color, contrast, or other local spatial features attract attention. However, subjects always gazed upon the image's main subject(s): Gaze position samples tended to occur on objects which belong to animal faces, human faces, or other subjects which represent the gist of the image. The background, such as sky and ground, always attracted the least attention.

Yet, despite these similarities, the saliency maps and importance maps don't always agree. Although we employed a relatively long viewing period, the saliency maps never yielded an object-level segregation which is enforced in the importance maps. For example, whenever a face occurred in an image, whether an animal face or a human face, the subjects' gaze positions always occurred on the face. Furthermore, as demonstrated by the bottom-most image in the left-hand group in Figure 4.3.1, which contains people in the background (the people are located toward the top-left corner of the image), the

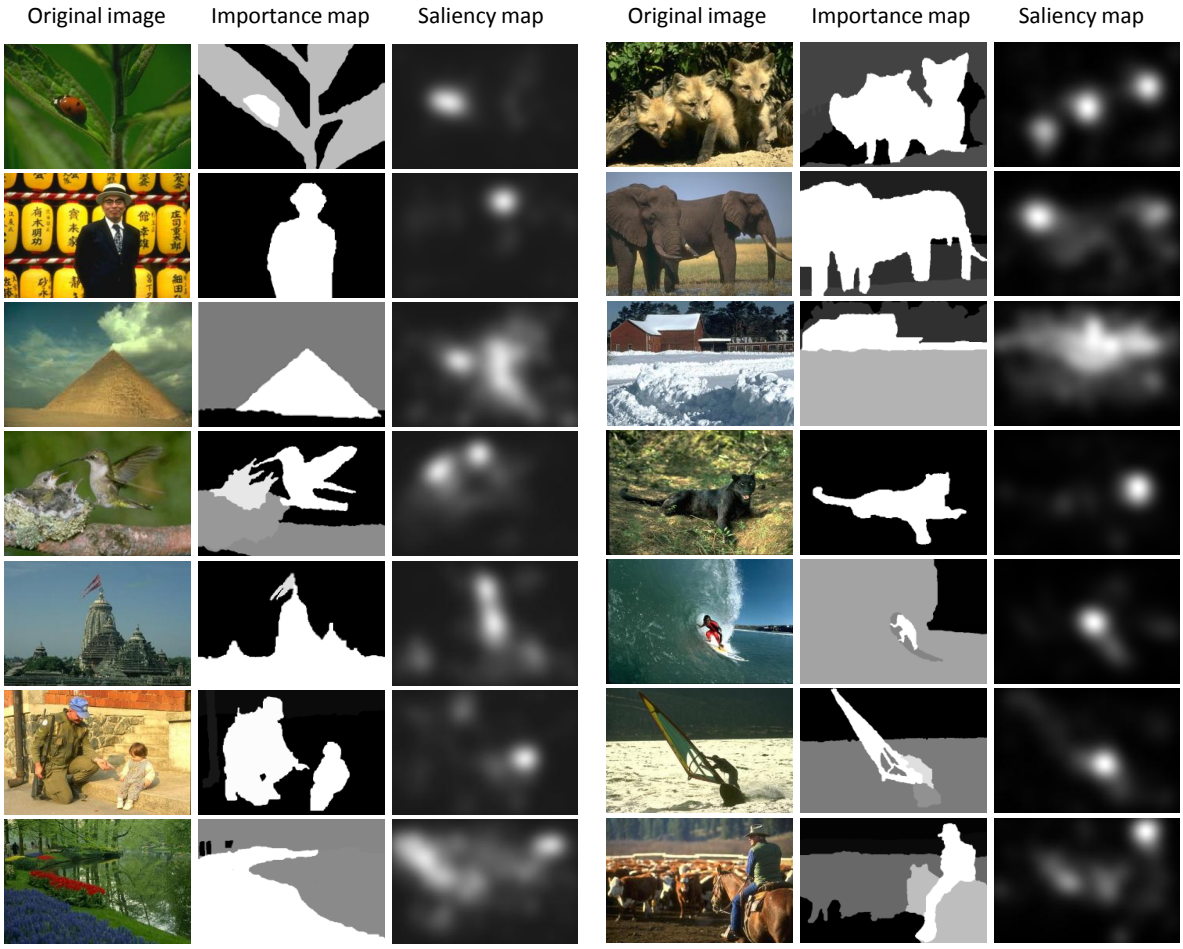


Figure 4.3.1: Representative results from the experiments. Note that the saliency maps are created based on the data of 15-second viewing.

importance ratings seem to be influenced by the artistic intent (e.g., the flowers and reflecting pond), whereas the saliency maps highlight only some of these regions.

4.3.2 Predicting the Main Subject, Secondary Objects, and the Background

The results of the qualitative analysis suggest a relationship between saliency maps and importance maps. One way to quantify this relationship is to attempt to predict the importance maps from the saliency maps using the object-level segmentations as side-information. This approach decouples the errors in prediction from the errors due to segmentation, since the latter would otherwise give rise to an artificially low measure of dependency between the two maps.

To predict the importance maps from the saliency maps (given the segmentations), the following two approaches were tested:

- *Mean Saliency*: For each object, we summed those values of the saliency map which occurred within the object, and then we divided this value by the total number of pixels in the object. For each image, the resulting set of per-object saliency values was then normalized to span the range $[0, 1]$.
- *Coverage Saliency*: For each object, we summed those values of the saliency map which occurred within the object, and then we divided this value by the number of pixels in the object that were gazed upon (specifically, the number of pixels that were covered by the fovea). For each image, the resulting set of per-object coverage saliency values was then normalized to span the range $[0, 1]$.

To facilitate the prediction, each importance map was quantized into three classes based on the importance values:

- *Main subjects*, which consisted of objects which received an importance value ranging from 2/3 to 1;
- *Secondary objects*, which received an importance value ranging from 1/3 to 2/3;
- *Background objects*, which received an importance value ranging from 0 to 1/3.

The results of the prediction are provided in Table 4.2 in the form of confusion matrices. Each row of each matrix represents the actual importance class, and each column represents the predicted class. An ideal prediction would yield a diagonal matrix with 100% values. As shown in Table 4.2(a), the average saliency can successfully predict the main subject approximately 81% of the time. Similarly, the background is successfully predicted approximately 47% of the time.[2] Coverage saliency, shown in Table 4.2(b), yields worse performance for main subjects, but slightly better performance for background objects.

[2]Note, however, that some objects contain only two levels of importance: main subject and background.

		Predicted		
		Main subject	Secondary subject	Background
Actual	Main subject	**80.5%**	29.8%	12.6%
	Secondary subject	12.5%	**42.6%**	40.7%
	Background	7.1%	27.6%	**46.7%**

(a)

		Predicted		
		Main subject	Secondary subject	Background
Actual	Main subject	**56.5%**	38.6%	8.2%
	Secondary subject	13.0%	**40.4%**	24.7%
	Background	30.5%	21.1%	**67.1%**

(b)

Table 4.2: Confusion matrices for predicting each object's importance from the gaze data.

4.3.3 Temporal Analysis I: Early vs. Later Gaze Positions

During normal viewing, because visual attention shifts from one object to another, the number of gaze position samples which occur on each object varies over time. For each of the three levels of importance (main subjects, secondary objects, background), we analyzed this time dependence. Specifically, we computed the number of gaze position samples per importance class which occurred within each 100-ms interval during the 15-second viewing time. The resulting three time curves, summed across all observers, are shown in Figure 4.3.2.

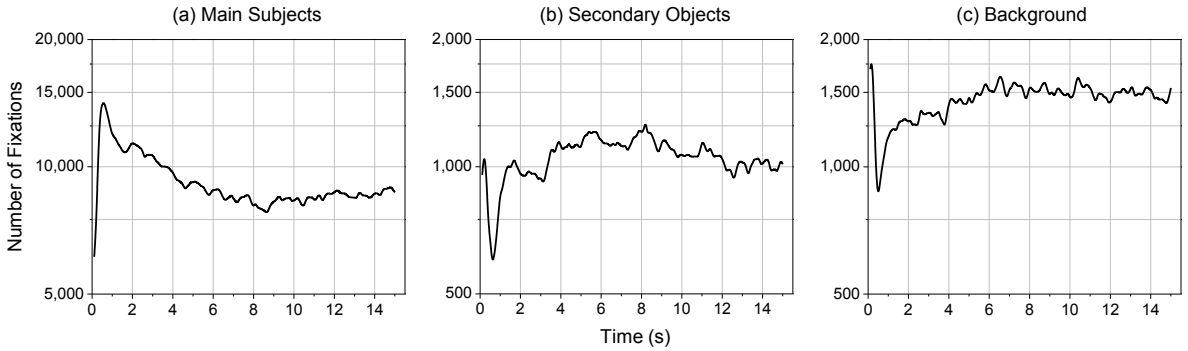


Figure 4.3.2: Total number of gaze position samples in (a) main subjects, (b) secondary objects, and (c) background objects computed in each 100-ms interval of the 15-second viewing time. Note that the scale for the vertical axis in the first graph is 10x that of the other two graphs.

The plots in Figure 4.3.2 clearly indicate that, on average, objects from different im-

portance classes attract considerably different amounts of visual attention. Specifically, throughout the 15-second viewing time, the main subjects always received the greatest number of gaze position samples—approximately 7-8 times greater than the number of samples for secondary and background objects.

Within 0-500 ms, the number of gaze position samples for the main subjects was already 4-6 times greater than the number of samples for secondary and background objects. This observation suggests bottom-up mechanisms can be effective at locating the main subjects in these images; this might result from the fact that photographers tend to increase the saliency of the main subjects via, e.g., retouching, selective focusing, or other photographic techniques. Between 500-2000 ms, there was a pronounced increase in the number of gaze position samples for the main subjects, while the number for the other two importance classes decreased in this period. These changes potentially indicate the influence of top-down mechanisms which might force observers to attend to the main subjects. After this process, the number of gaze position samples for the main subjects slightly decreased and those for the other two classes slightly increased. This latter change may imply that the observers attempt to explore the whole image, but their attention is still held by the main subjects.

These three time curves suggest that the relationship between visual salience and visual importance may be time dependent. In particular, the fact that the main subjects attract the most attention within 0-2000 ms suggests that these early gaze position samples might be a better predictor of visual importance for the main subjects than previously achieved using all samples. Accordingly, we predicted the importance maps by using the samples taken from only the first 0-2000 ms. Table 4.3 lists the resulting confusion matrix computed (using Mean Saliency method) based on gaze data of the first 2 seconds. Figure 4.3.3 depict representative importance maps predicted from the gaze data taken from all 15 seconds and from only the first two seconds. By using only these early gaze data, better prediction is achieved for the main subjects.

		Predicted		
		Main subject	Secondary subject	Background
Actual	Main subject	**89.0%**	43.5%	12.4%
	Secondary subject	3.3%	**43.5%**	27.2%
	Background	7.7%	13.0%	**60.5%**

Table 4.3: Confusion matrix for predicting importance from the first 2 seconds of gaze samples using mean saliency.

4.3.4 Temporal Analysis II: Normalized Scanpath Saliency

To quantify how well the visual saliency fits the importance map at every moment, we used normalized scanpath saliency (NSS), which has previously been used for evaluating

Figure 4.3.3: Representative results of using all gaze samples vs. only those from the first two seconds to predict the importance maps.

the performance of visual attention computational models [Pang 08]. We consider the importance map as some kind of "predicted saliency map", and compare it with gaze points .

Let $R_n(t)$ be a set of all pixels in the circular region centered at the eye positions of test subject n. This region is with a radius of 14 pixels which is computed based on the viewing condition. Then, the NSS value computed for each 100-ms time-slice t is defined as

$$NSS(t) = \frac{1}{N_i}\sum_{i=1}^{N_i}\left(\frac{1}{N_s}\sum_{n=1}^{N_s}\frac{1}{\sigma(p(x))}\left\{\max_{x(t)\in R_n(t)} p(x(t)) - \bar{p}(x)\right\}\right),$$

where N_i is the number of images, N_s is the total number of test subjects, $\bar{p}(x)$ and $\sigma(p(x))$ are the mean and the variance of the model's output density map, respectively.

An NSS value of unity indicates the subjects' eye positions fall on an object whose importance value is one standard deviation above average. A low NSS value indicates a lesser agreement between the eye positions and objects' importances. Figure 4.3.4(a) shows the NSS value computed from the saliency map generated from the gaze points within each 100-ms time-slice.

4.3.5 Temporal Analysis III: Kullback-Leibler Divergence

The Kullback-Leibler divergence is usually used to compute the degree of dissimilarity between two probability density functions, which are deduced from the human fixation density map and the predicted saliency map. Here, we used it to compare each saliency map with its corresponding importance map, interpreting each as a pdf. The Kullback-Leibler divergence is given by

$$KL(p|h) = \sum_x p(x)\log\frac{p(x)}{h(x)},$$

where h is the human importance maps, and p is the fixation pattern smoothed by a Gaussian kernel. $KL = 0$ means that the probability densities are strictly equal. The KL divergence computed from the saliency map generated from the gaze points within each 100-ms time-slice is shown in Figure 4.3.4(b).

4.3.6 Temporal Analysis IV: Linear Correlation Coefficient

The linear correlation coefficient (CC) measures the strength of a linear relationship between two variables. It is also a common measure to evaluate the performance of computational models of visual attention. CC has a value between -1 and +1. If the correlation is close to +1 or -1, then there is an almost perfectly linear relationship between the two variables, whereas a value of zero indicates no linear relationship. Here, we

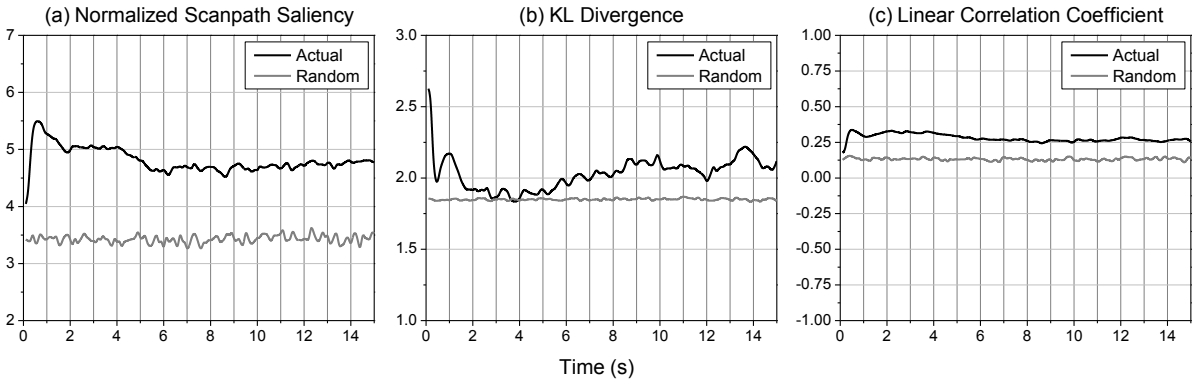


Figure 4.3.4: (a) Normalized scanpath saliency, (b) KL divergence, and (c) linear correlation coefficient, each computed between the importance map and the saliency map generated from the gaze points within each 100-ms time-slice, averaged across all subjects. The black line in each plot corresponds to values computed using the actual gaze positions measured in Experiment II; the gray line in each plot corresponds to values computed using random gaze positions.

calculate CC to compare the fixation patterns and the importance map. The correlation coefficient for each 100-ms time-slice, $\rho(t)$, is given by

$$\rho(t) = \frac{\sum_{x,y} (I(x,y) - \mu_I)(S(x,y) - \mu_S)}{\sigma_I \sigma_S},$$

where I denotes the importance map, S denotes the saliency map computed in the 100-ms time interval, and where $\mu_I, \mu_S, \sigma_I, \sigma_S$ are the respective means and standard deviations of these maps. $\rho = 0$ indicates that there is no correlation. The linear correlation coefficient computed from the saliency map generated from the gaze points within each 100-ms time-slice is shown in Figure 4.3.4(c).

4.3.7 Temporal Analysis V: Volume Under the ROC Surface

An receiver-operating-characteristic (ROC) plot allows a classifier to be evaluated and optimized over all possible operating points. The area under the ROC curve (AUC) is a usual performance evaluation criterion in two-class pattern recognition problems. An extension of AUC to the multiclass case is the volume under the ROC surface (VUS). Here, we use VUS to evaluate the performance of predicting the importance class from the saliency maps computed for each time-slice.

Landgrebe et al. [Landgrebe 06] proposed a simplified VUS measure, which ignores specific intraclass dimensions, and regards only interclass performances (the diagnoal entries in the confusion matrix). It was found that some classifiers compete in terms of error-rate, but have significantly different VUS scores, illustrating the importance of the

VUS approach. Here, the simplified ROC dimensionality is three, since we have three importance classes.

We calculate the VUS by this simplified measure for each 100-ms time-slice. For the prediction, we used three methods: Mean saliency, sum saliency, and coverage saliency. (Sum saliency is the sum of the saliency-map values which occurred within the object; it is the mean saliency for an object multiplied by the number of pixels for that object.) The results of this analysis are shown in Figure 4.3.5.

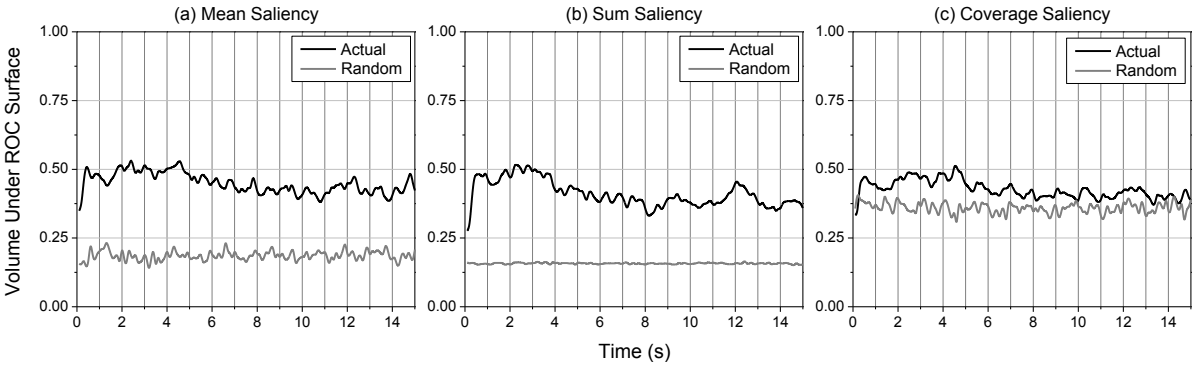


Figure 4.3.5: Volume under the ROC surface for the classification task of predicting primary-, secondary-, and non-ROI from the saliency map generated from the gaze points within each 100-ms time-slice, averaged across all subjects. The saliency maps were computed by using (a) mean saliency, (b) sum saliency, and (c) coverage saliency. The black line in each plot corresponds to values computed using the actual gaze positions measured in Experiment II; the gray line in each plot corresponds to values computed using random gaze positions.

4.3.8 Temporal Analysis VI: Vector Distance

The distance between two vectors might be used as another method to evaluate the prediction performance. The importance values of all the objects in the same image are several scalars, which can compose a vector. Hence, there exist two vectors for one image. One is generated by the predicted importance values, and the other one is generated by the human importance values. Measuring the distance between these two vectors could be an approach to quantify the accuracy of prediction. The distance is given by

$$d(t) = \frac{1}{N_i} \sum_{i=1}^{N_i} \left\| \mathbf{v}_{human,i} - \mathbf{v}_{predict,i} \right\|_2 ,$$

where $N_i = 80$ denotes the number of images. The variable $\mathbf{v}_{human,i}$ denotes the (ground-truth) vector of importance values for image i measured in Experiment I. The variable

$\mathbf{v}_{predict,i}$ denotes the predicted vector of importance values for image i, where the prediction was made by using the saliency map generated from the gaze points within each 100-ms time-slice. $\|\cdot\|_2$ denotes the vector euclidean norm. The results of this analysis are shown in Figure 4.3.6.

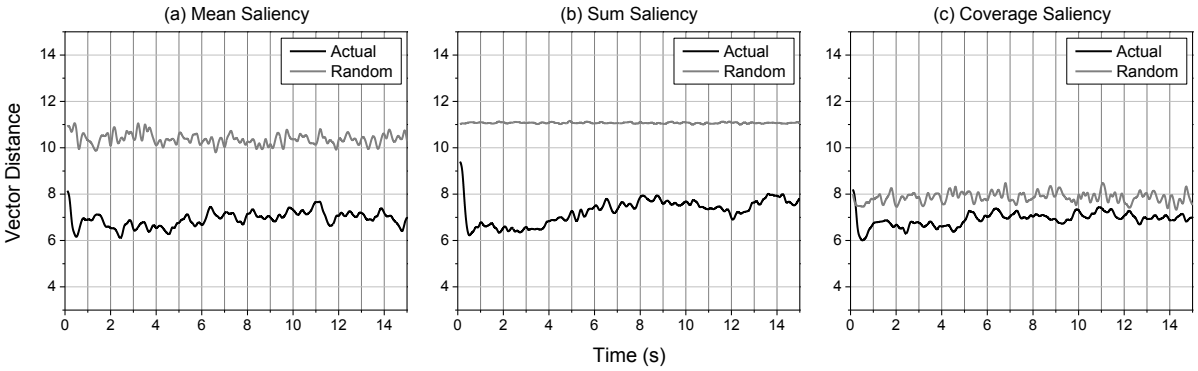


Figure 4.3.6: Vector distance between the ground-truth and predicted importance values generated from the gaze points within each 100-ms time-slice, averaged across all subjects. The saliency maps were computed by using (a) mean saliency, (b) sum saliency, and (c) coverage saliency. The black line in each plot corresponds to values computed using the actual gaze positions measured in Experiment II; the gray line in each plot corresponds to values computed using random gaze positions.

4.4 Discussion

In this study, we have attempted to quantify the similarities and differences between bottom-up visual salience and top-down visual importance. The implications of these initial findings for image processing are quite important. Several algorithms have been published which can successfully predict gaze patterns (e.g., Refs. [Itti 98, Le Meur 06]). Our results suggest that these predicted patterns can be used to predict importance maps when coupled with a segmentation scheme. In turn, the importance maps can then be used to perform importance-based processing such as auto-cropping, enhancement, compression, unequal error protection, and quality assessment.

Yet, as we have seen here, even if we use ground truth visual saliency map, the predictions are not as great as one might have expected. Below we describe some possible explanations along with some implications for visual coding.

4.4.1 The Segmentation Problem

When the observers look at an interesting object, they do not fixate upon all the areas of the object. Only selected regions of an object attract the vast majority of visual

fixations. The clearest examples of these are for human/animal faces vs. the rest of their bodies; for most images, people looked only at the faces. Figure 4.4.1 depicts some prime examples of this observation.

Figure 4.4.1: Examples of subjects visually attending to only select regions of objects, most notable for images containing faces.

Despite the fact that only select regions of an object attract the vast majority of visual fixations, these objects were rated by subjects to be of high importance. This suggests that either: (1) humans tend to base the importance of the entire object on the importance of the most interesting region (max operator under visual summation); or (2) some objects could benefit from another level of segmentation. In Experiment I, subjects were forced to rate object-level importance, where the object was defined by the given segmentation. One could possibly obtain ratings for each object's sub-objects and draw a different conclusion. Although the segmentations used in Experiment I were based on hand-segmentations provided by the Berkeley dataset, the results of our analyses are restricted by these segmentations.

4.4.2 The Border-Ownership Problem

When a gaze samples falls on or near the border of an object, there remains a question as to which object the gaze sample should be assigned. Indeed, perception itself dictates which object owns the border (the "border ownership" problem; Ref. [Nakayama 89]).

For example, as shown in Figure 4.4.2(a), the gaze position samples tend to suggest that the subjects are visually attending to the whale, despite the fact that some of the near-border gaze samples actually fall upon the background (water). However, as shown in Figure 4.4.2(b), it is not always easy to decipher to which object the subjects were visually attending. In this latter case, it is not clear whether near-border samples should be assigned to the men or the elephants.

Because of this ambiguity, the importance maps for images such as that depicted in Figure 4.4.2(a) may reflect abnormally high values of importance for the background. In terms of border ownership, the region to which the border is assigned is perceived as the figure. However, without knowledge of each subject's instantaneous figure/ground

Figure 4.4.2: Examples of ambiguities regarding to which object a gaze position sample should be assigned.

interpretation, there is no correct way to assign a gaze position sample to a particular object.

4.4.3 Bottom-Up vs. Top-Down Visual Coding

Two mechanisms are at work when humans look at an image: bottom-up and top-down. The results of our temporal analyses suggest that visual attention varies over time and that the extent to which visual attention and visual importance maps are related also varies with time.

The normalized scanpath saliency, KL divergence, and linear correlation coefficient all point to the fact that the relationship between visual attention and visual importance is strongest in the first two seconds of the 15-second observation interval, which implies that top-down mechanisms dominate eye movements during this period. This finding was also observed from the prediction-based analyses. A good prediction for the primary ROI implies that top-down mechanisms dominate eye movements (observers are looking at the most important objects), whereas a bad prediction suggests either that bottom-up mechanisms are at work, or that the two mechanisms are competing.

These findings suggest a possible strategy for human visual coding. If the human visual system can so rapidly identify the main subject(s) in a scene, such information can be used to prime lower-level neurons to better encode the visual input. This strategy may play an important role toward improving neural coding efficiency or facilitating higher-level tasks such as scene categorization.

Several researchers have suggested that rapid visual priming might be achieved via

feedback and/or lateral interactions between groups of neurons after the "gist" of the scene is determined [Friedman 79, Oliva 05]. The results of this study provide psychophysical evidence that lends support to a gist-based strategy and a possible role for the feedback connections that are so prevalent in mammalian visual systems.

4.5 Conclusions

This chapter presented the results of two psychophysical experiments and an associated computational analysis designed to quantify the relationship between visual salience and visual importance. We found that saliency maps and importance maps are related, but perhaps less than one might expect. The saliency maps were shown to be effective at predicting the main subjects. However, the saliency maps were less effective at predicting the objects of secondary importance and the unimportant objects. We also found that the vast majority of early gaze position samples (0-2000 ms) were made on the main subjects. This suggests that a possible strategy of the human visual system is to quickly locate the main subject(s) in the scene.

Key points

Context

- ❏ Visual salience and visual importance come from the two different mechanisms of visual attention, the bottom-up mechanism and the top-down mechanisms, respectively. They provide important insights into how biological visual systems address the image-analysis problem.
- ❏ However, despite the differences in the way (bottom-up) visual salience and (top-down) visual importance are determined in terms of human visual processing, both salience and importance have traditionally been considered synonymous in the signal-processing community: they are both believed to denote the most visually "relevant" parts of the scene.

Contributions

- ❏ We conducted two psychophysical experiments and an associated computational analysis designed to quantify the relationship between visual salience and visual importance.
- ❏ In the first experiment, importance maps were collected for a database of images by asking human subjects to rate the relative visual importance of each object within hand-segmented images.
- ❏ In the second experiment, experimental saliency maps were computed from visual gaze patterns measured for these same images by using an eye-tracker and task-free viewing.
- ❏ By comparing the importance maps with the saliency maps, we found that the maps are related, but perhaps less than one might expect. When coupled with the segmentation information, the saliency maps were shown to be effective at predicting the main subjects. However, the saliency maps were less effective at predicting the objects of secondary importance and the unimportant objects.
- ❏ We also found that the vast majority of early gaze position samples (0-2000 ms) were made on the main subjects, suggesting that a possible strategy of early visual coding might be to quickly locate the main subject(s) in the scene.

Chapter 5

Eye-tracking database for stereoscopic 3D natural-content images

The availability of ground truth, i.e. eye-tracking data, is always of particular importance, especially in computational modeling of visual attention. Publicly available eye-tracking databases can facilitate fair comparisons amongst computational models. Therefore, several image and video eye-tracking databases have been published in the community.

However, most of the databases publicly available in recent years are created for 2D images or videos. Nowadays, studies related to stereoscopic 3D have been recently gaining an increasing amount of attention, but any eye-tracking database for 3D content is still missing in the community. The lack of ground truth leads to the difficulties of quantitatively assessing and comparing the performance for most of the existing 3D computational models, as well as the influence of depth features.

From this chapter, our work starts to focus on the topics related to the visual attention in stereoscopic 3D viewing condition. For simplicity of notation, we would like to use the term "3D" to refer to "stereoscopic 3D" in the remainder of this article. We start our work on 3D visual attention modeling by an attempt to firstly solve the problem of the lack of ground truth. In this chapter, we introduce a binocular eye-tracking experiment based on stereoscopic 3D images. Based on this work, we create and publish a new eye-tracking database containing eighteen stereoscopic natural-content images, the corresponding disparity maps, and eye movement data of both eyes.

5.1 Introduction

Three-dimensional content increases the sensation of presence through the enhancement of depth perception. To achieve this task, binocular depth cues, such as binocular disparity, are introduced and fused together with other monocular depth cues in an adaptive way. Studies related to stereoscopic 3D have been recently gaining an increasing

amount of attention because of the emergence of 3D content (in cinema and home) and the recent availability of high-definition 3D-capable acquisition and display equipments.

The very first challenge that needs to be addressed is how to reliably collect ground truth data. However, in the community of computational visual attention modeling, the lack of ground truth is having a negative impact on the development of 3D visual attention models. In order to provide reliable ground truth to the community for evaluating the performance of the models, we conduct an eye-tracking experiment for creating a new FDM database for 3D still images.

Due to the difference between the viewing conditions of 2D and 3D, the eye-tracking experiments for 3D content might considerably differ from the eye-tracking experiments for 2D content. Moreover, adapting traditional eye-tracker for the task of recording eye movements during viewing 3D content also faces some new challenges.

- Binocular eye-trackers are necessary. Most previous studies on 2D visual attention used monocular eye-trackers, or binocular eye-trackers with the option of using the data from only one eye [Huynh-Thu 11a]. These studies assumed a similar behavior for each eye during the viewing of 2D content. Due to the disparity between the left and the right view in 3D content, the difference of gaze positions between the two eyes may prove to be more critical. It is thus necessary for the eye-tracker to be with the capacity of recording individually the gaze positions of the two eyes.

- Since eye-tracking is performed binocularly, and the difference of two eyes' gaze points is relevant, the calibration procedure for an experiment using 3D stimuli is now facing new challenges. During the calibration, the reliabilities of the data recorded from both eyes need to be verified individually. This demand make the calibration step more difficult and time-consuming as compared to a 2D eye-tracking experiment.

- During the viewing of 3D content using glasses (including active shutter glasses and passive polarized glasses), the luminance of the stimuli when watched through glasses has to be verified before the experiment. Moreover, the luminance of the environment also needs to be adjusted accordingly. These controls of luminance are to make sure the pupil is with appropriate size which enables the eye-tracker to detect the pupil position and the corneal reflection (i.e. the first Purkinje image).

- So far, it is still difficult to totally get rid of the visual discomfort and visual fatigue during the viewing of stereoscopic 3D content [Hoffman 08]. These problems also have impact on eye-tracking experiments. Firstly, the issue of comfortable viewing zone has to be taken into account when selecting stimuli [Chen 10]. Secondly, the experimenter need to be cautions on the duration of experiment. As compared to eye-tracking experiments using 2D stimuli, it is of particular importance to have a larger number of rest and re-calibration for 3D viewing condition.

In addition to the issues introduced above, the relative smaller amount of 3D content and the difficulties in obtaining reliable ground truth of depth information also lead to the lack of 3D eye-tracking database for computational modeling of visual attention. In the following sections, we will introduce in detail our eye-tracking experiment, which provides a new eye-tracking database containing eighteen stereoscopic natural-content images, the corresponding disparity maps, and eye movement data of both eyes.

5.2 Stimuli

The stereoscopic images used in the proposed database were acquired from two sources: (1) the Middlebury 2005/2006 image dataset, and (2) the IVC 3D image dataset.

5.2.1 Image sources

The Middlebury 2005/2006 dataset

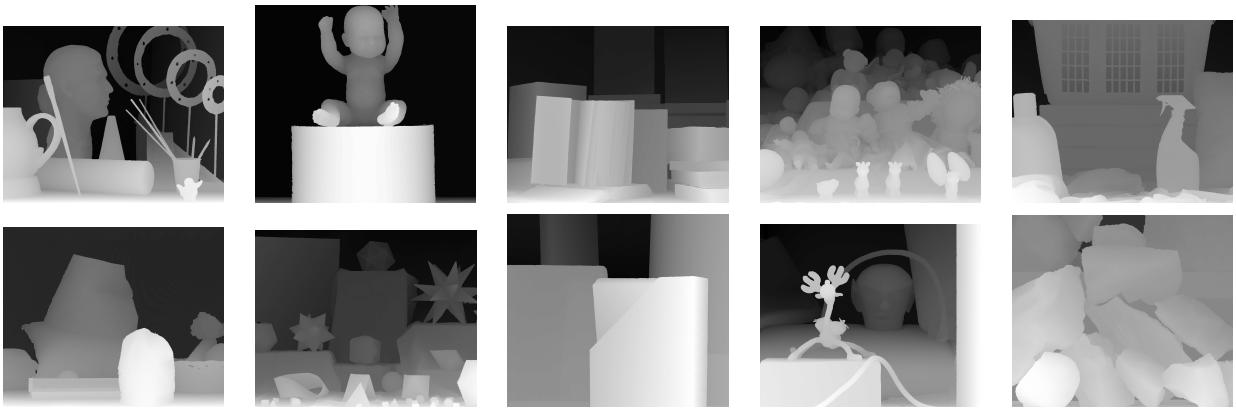
Scharstein et al. [Scharstein 07] created 30 multi-view 3D images. Each image corresponds to one particular indoor scene taken from a close-up view. Each of them consists of 7 rectified views taken from equidistant points along a line, as well as ground-truth disparity maps for view 2 and 6. In this image acquisition system, the focal length was set to 3740 pixels, and the directions of cameras were parallel. The ground-truth disparity maps were created by using an auto-mated version of the structured-lighting technique of [Scharstein 03]. The images are with a resolution about 1300×1000 pixels, and with about 150 different integer disparity values.

We selected 10 images from the Middlebury 2005/2006 image dataset for our eye-tracking experiment. Considering visual comfort, view 2 and view 4 were used as the left view and right view respectively for each scene. This selection is made to avoid the appearance of excessive relative disparity in one scene. The baseline between the two views was thus supposed to be 800 mm. The images selected are shown in Figure 5.2.1 with their corresponding disparity maps.

The IVC 3D image dataset

We produced a set of eight 3D videos by using Panasonic AG-3DA1 3D camera (see [Urvoy 12]). One frame from each video was selected by the authors to create this IVC 3D Image Dataset. Each video consists of two sequences representing the left view and the right view respectively. Both sequences were full-HD resolution (1920×1080 pixels). This set of videos contains two outdoor scenes and six indoor scenes, which were taken in University of Nantes. Compared to the Middlebury database, the scenes in this set of videos have a higher average depth value. The distance between camera and the first object in the scene is at least two meters. For color adjustment, a white-balance

(a)

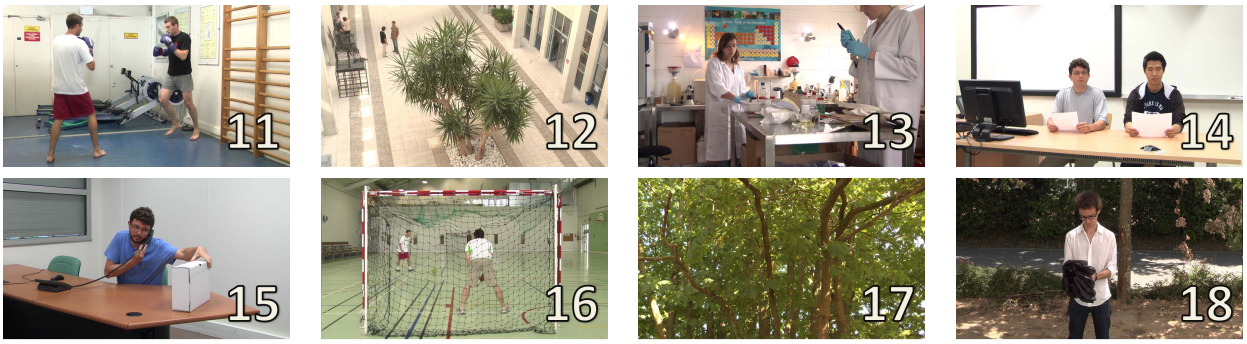


(b) The corresponding disparity maps

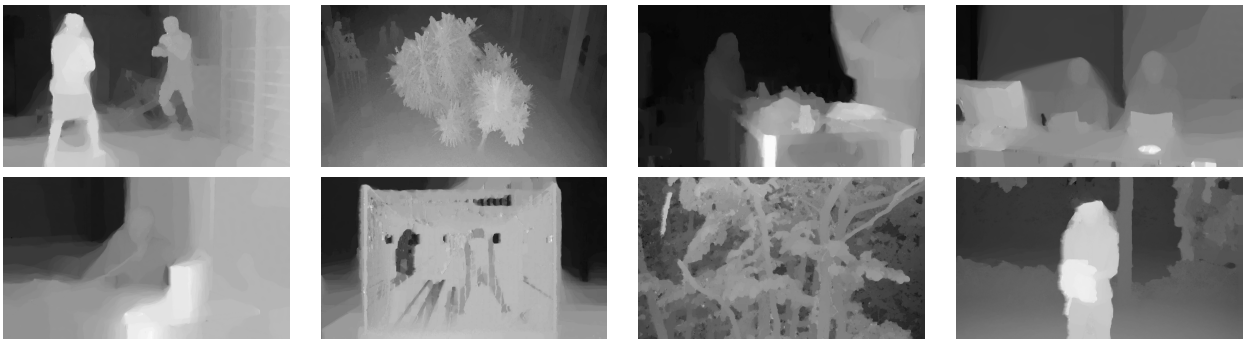
Figure 5.2.1: Illustration of the images obtained from the Middlebury dataset. (a) The original images. The number at the bottom right corner indicates the index of each image. (b) The corresponding disparity maps.

adjustment was done by using a white A4 paper at approximately 100 cm in front of the camera before each shooting.

Without the deployment of any depth range sensors during the acquisition of videos, depth maps of IVC 3D image database were obtained by a post-processing depth map estimation on the stereo-pair images. The depth map estimation we deployed was an optical flow approach proposed by Werlberger et al. [Werlberger 09, Werlberger 10]. The general idea of this approach was inspired by 2D motion estimation algorithms that take advantage of optical flow estimation. To create the ground truth disparity map, we computed the 'left-to-right' disparity map which represents the displacement of each pixel in the left view. Both the images and their corresponding disparity maps are showed in Figure 5.2.2.



(a)



(b)

Figure 5.2.2: Illustration of the images obtained from the IVC 3D image dataset. (a) The original images. The number at the bottom right corner indicates the index of each image. (b) The corresponding disparity maps.

5.2.2 Stimuli adjustment

Before being used in the eye-tracking experiment, the images obtained from the Middlebury 2005/2006 dataset need to be modified in order to get rid of the problem of "stereo

window violation" in our stereoscopic 3D display environment. Moreover, the problems existing in the depth maps needs also to be fixed. A post-processing step of the the data from the Middlebury dataset is thus of particular importance.

5.2.2.1 Stereo window violation removal

Since the cameras used for generating the views were directed in parallel, we can consider that they converged at an infinite point. This setup makes a direct utilization of the two images from view 2 and view 4 as stereo pair lead to a so-called "stereo window violation" [Halle 05]. All pixels in the scene were perceived as being in front of the screen plane, meanwhile, a certain area close to the left edge of the left view, and a certain area close to right edge of the right view, were displayed only in the left view and right view, respectively. Serious visual rivalry and visual discomfort could happen when these two areas were looked at.

Apart from visual rivalry, another problem is an insufficient exploitation of depth range. It was suggested that a 3D scene should be located in a limited depth range named comfortable viewing zone [Chen 10] ranging from the back to the front of the screen plane. This comfortable viewing zone is computed considering the capability of binocular images fusion, and conflicts among different depth cues such as accommodation, vergence [Hoffman 08] and blur. In our case, if the entire scene was displayed only in front of the screen plane, the depth range that could be taken advantage of would be thus limited to approximate a half.

To overcome these problems, we adjusted the depth range of the scene by using the method proposed by Chamaret et al. [Chamaret 10]. We shifted the left view towards left, the right view towards right. This shifting of two views in opposite directions equals to adding a constant negative disparity for every pixel in the two views. The amount of added disparity was calculated by:

$$D_{add} = \frac{D_{\min} - D_{\max}}{2} \tag{5.2.1}$$

where $D_{\min}$ and $D_{\max}$ denote the minimum and maximum disparity value in the scene, respectively. Therefore, half of the depth range of the scene was moved to the back of the screen plane, while the other half was still in front of the screen plane.

5.2.2.2 Disparity map refining

Although most of the areas in the disparity maps of the images provided by the Middlebury dataset were with high accuracy, the disparity values were still unknown at some locations, such as some deep holes surrounded by several objects and the edges where occlusion happened (as shown in Figure 5.2.3).

For the first case, the area of these regions was usually large, and the actual disparity value was different from all surrounding. We thus did a manual refining by justifying

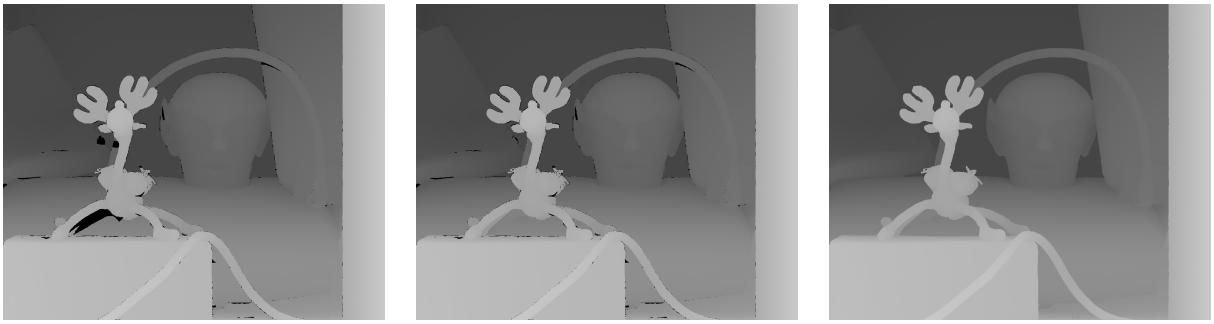


Figure 5.2.3: Example of disparity map refining.

the depth value of these regions considering the content of the whole scene. We first checked if a region was part of background or any objects, both of which had reliable depth information at some other locations in the scene, then we manually assigned the same depth value to these regions.

For the second case, the region with unknown disparity value in the disparity map usually consisted of some (groups of) pixels which covered small areas and located sparsely along the edges. The disparity values of these pixels were close or even equal to surrounding pixels. An automatic refining was thus performed by using an inpainting algorithm proposed by Criminisi et al. [Criminisi 03], which was an exemplar-based inpainting algorithm which fills holes in a visually plausible way and persists one-dimensional patterns, such as lines and object contours.

5.3 Apparatus and procedures

Stimuli were displayed on a 26-inch (552×323 mm) Panasonic BT-3DL2550 LCD screen, which has a resolution of 1920×1200 pixels, and the refresh rate was 60 Hz. Each screen pixel subtended 61.99 arcsec at a 93 cm viewing distance. The maximum luminance of the display was 180 cd/m^2, which yielded a maximum luminance of about 60 cd/m^2 when watched through glasses. Observers viewed the stereoscopic stimuli through a pair of passive polarized glasses at a distance of 93 cm. The environment luminance was adjusted according to each observer, in order to let the pupil have an appropriate size for eye-tracking. SMI RED 500 remote eye-tracker was used to record the eye movements. A chin-rest was used to stabilize observer's head.

The eighteen stereoscopic images (ten from the Middlebury dataset and eight from IVC 3D image dataset) as well as their 2D version (containing two copies of the left view images) were presented in a random order. Presentation time of each scene was 15 seconds. Between every two scenes, a center point was showed for 500 ms at the screen center with zero disparity. Subjects were required to do a free-viewing of the scene. A nine-point calibration was performed at the beginning of the experiment, and repeated

every ten scenes. The quality of calibration was verified by the experimenter on another monitor. Participants could require for a rest before every calibration started. Each observer was required to have at least three rests during the whole observation. All the experiments were conducted from 10:00 to 12:00 and from 14:00 to 16:00, in order to decrease the feeling of fatigue as much as possible.

5.4 Participants

Thirty-five subjects participated in the experiment, including 28 students, 3 engineers, 1 university staff, and 3 people without job. Note that none of the subjects in this group ever participated in the experiment for probability distribution modeling. Subjects ranged in age from 18 to 46 years old. The mean age of subjects was 24.2 years old. All the subjects had either normal or corrected-to-normal visual acuity, which was verified by three pretests before the start of eye-tracking experiment:

1. Monoyer chart test was performed to check the acuity (subject must get results higher than 9/10);
2. Ishihara test was performed to check color vision (subject should be without any color troubles);
3. Randot stereo test was performed to check the 3D acuity (subject should get results higher than 7/10).

5.5 Fixation density map creation

In order to take into account both the position and duration of eye movements, all gaze points recorded by the eye-tracker from both the left eye and the right eye were used to create the fixation density maps. The gaze points maps from each eye were first created respectively. The left gaze points map was created by directly using the coordinates of gaze positions of the left eye. However, according to the argument that it would be compelling, in a biological sense, to accommodate shifts in the position of an attended event from one eye to another [Huynh-Thu 11a], we created the right gaze points map by adding a displacement, horizontally and vertically, on the coordinates of each right-eye gaze point. The displacements of each gazed point were obtained from the 'right-to-left' disparity map computed by the same approach as the one used to create the ground truth disparity maps of IVC 3D images database.

The two gaze points maps were then summed, and filtered by a two-dimensional Gaussian kernel to account for 1) the decrease in the visual accuracy with increasing eccentricity from the fovea, and 2) the decrease in the accuracy of the eye tracker. The standard deviation of the Gaussian kernel used in our creation of saliency maps was equal to 2 degrees of visual angle.

5.6 Conclusion and perspectives

In this chapter, we introduce a binocular eye-tracking experiment based on stereoscopic 3D images. Based on this work, we create and publish a new eye-tracking database containing eighteen stereoscopic natural-content images, the corresponding disparity maps, and eye movement data (including fixation density maps) of both eyes. The created FDM can be used, in the future, for evaluating the performance of 3D visual attention models.

Nevertheless, when comparing with the amount of eye-tracking databases for 2D image and videos, an eye-tracking database containing only eighteen images is far from enough. Therefore, more efforts need to be dedicated in this issue in the future work. An eye-tracking database containing a larger number of images, or 3D videos, will make further contribution to the community.

Furthermore, concerning the collection of eye-tracking data in the context of 3D-TV, there are still several challenges. For instance, one of the challenge is that additional efforts have to be put into the calibration procedure. It has been proposed that a "volumetric calibration" is required in an eye-tracking experiment using 3D stimuli, i.e. showing points at different locations and different depth planes [Huynh-Thu 11a]. Another challenge comes from the interpretation of eye-tracking data. Current eye-tracking equipment can only provide a two-dimensional spatial gaze location individually for each eye, and the resulting fixation density maps are also two-dimensional topographical maps. We argue that a three-dimensional fixation density map, which might be obtained based on the triangulation of corresponding gaze points from two eyes and shows thus the 3D location of gaze points, would be more appropriate for the studies of 3D visual attention. However, this new type of saliency map requires eye-trackers with higher level of accuracy, which thus raises another challenge.

Key points

Context

- ❏ Studies regarding computational modeling of 3D visual attention have been gaining an increasing amount of attention. The amount of demand for ground truth is also increasing.
- ❏ Due to the difference between the viewing conditions of 2D and 3D, the eye-tracking experiments for 3D content considerably differ from the eye-tracking experiments for 2D content. These difference increase the difficulty of creating eye-tracking databases for 3D content.
- ❏ So far, any public eye-tracking databases for 3D content can be hardly found in the community. The lack of ground truth leads to the difficulties of quantitatively assessing and comparing the performance of 3D visual attention models

Contributions

- ❏ We conduct a binocular eye-tracking experiment, then we create a new eye-tracking database containing eighteen 3D natural-content images (as well as their corresponding 2D version), the corresponding disparity maps, and eye movement data (including fixation density maps) of both eyes. This database helps in solving the problem of lacking ground truth in the research area of 3D visual attention modeling.

Computational Modeling of Visual Attention for Stereoscopic 3D Contents and prospective application

"The function of the human eye ... was described by a large number of authors in a certain way. But I found it to be completely different.."

(Leonardo Da Vinci)

Chapter 6

Influence of depth on stereoscopic 3D visual attention: introducing depth-bias

From the previous chapter, we have shifted our attention from 2D content to 3D content. Our work is now focusing on the computational modeling of 3D visual attention. As we know, the most important factor that distinguish the 3D viewing experience from 2D is the enhancement of depth perception. It is thus of particular importance to examine how the added depth perception in 3D viewing condition affects the deployment of visual attention.

In this chapter, we firstly introduce the background and previous studies on investigating the viewing behavior for stereoscopic 3D viewing condition. Secondly, we introduce a binocular eye-tracking experiment using synthetic stimuli, as well as the associated analyses, in order to study a so-called "depth-bias" which might help in computational modeling of 3D visual attention. Finally, the sections of discussion and conclusion are presented at the end of this chapter.

6.1 Introduction

In the studies regarding to the deployment of visual attention on planer screen, it has been found that observers' fixations exhibit a marked bias towards certain areas on the screen. In the viewing of 2D images or videos, a so-called "center-bias" (or "central fixation bias") has been demonstrated: gaze fixations are biased towards the center of the scene [Tseng 09, Tatler 07]. However, in the viewing of 3D content (on planar stereoscopic displays), depth perception is enhanced by binocular depth cues (e.g. binocular disparity [Neri 04, Howard 95, Wheatstone 38]). The viewing behavior of observers is thus largely changed due to the variation of depth perception. It has been recently shown that observers' fixations are biased not only towards the center area on the screen but also towards certain depth planes in the scene [Jansen 09, Wang 11b, Ramasamy 09]. It is thus reasonable to suppose the existence of a so-called "depth-bias".

In the area of developing computational models of visual attention in stereoscopic visualization, some hypotheses of depth-bias have been proposed (see [Maki 96b, Maki 00, Zhang 10, Chamaret 10]). Several of them consist of a similar architecture: saliency is computed by using 2D visual features and is then weighted according to depth information. Most of these studies assumed that areas or objects close to the viewer were more salient than distant ones. However, psychophysical studies of investigating this depth-bias on planar stereoscopic display are still lacking.

6.2 Previous studies

One of the few studies comes from Jansen et al. [Jansen 09] who investigated the viewing behavior in the observation of 2D and 3D still image. They conducted a free-viewing task on 2D and 3D version of the same set of images with natural content, pink noise or white noise. They found that viewer fixated closer locations earlier than more distant locations in both 3D images and 2D images. This result is also consistent with the works of Wang et al. [Wang 11b], who found that the closest object in a scene always attracted most fixations. Ramasamy et al. [Ramasamy 09] showed that observers' gaze points are more concentrated at the far end (in terms of depth) when viewing a scene containing long deep hallway. This result is inconsistent with the Jansen et al.'s result.

The inconsistency between the conclusions of these two studies might be due to the stimuli they used in their experiments. They used images with natural content which contain many visual features other than depth (e.g. color, intensity contrast, orientation, center-bias). These features might affect the deployment of observers' visual attention in both bottom-up way and top-down way [Wolfe 04, Itti 98, Le Meur 06, Bruce 09, Wang 10]. Therefore, it is important to get rid of the influence of 2D visual features on visual attention in order to investigate only the effect of the depth-bias. Using synthetic stimuli which are properly designed and controlled may be helpful to avoid any other side effects beside the bias under study as opposed to less controlled natural stimuli.

6.3 Experiment

We conducted a binocular eye-tracking experiment by showing synthetic stimuli on a stereoscopic display. Observers were required to do a free-viewing task. Gaze positions of both eyes were recorded, and both the location and the depth of fixations were computed. Stimuli presented during this experiment were designed in such a way that depth would affect eye movements independently from other visual features.

6.3.1 Participants

Twenty-seven subjects participated in this experiment (12 males and 15 females). The subjects ranged in age from 18 to 44 years. The mean age of the subjects was 22.8 years old. All of them were naive to the purpose of the experiment, and were compensated for their participation in the experiment. All the subjects had either normal or corrected-to-normal visual acuity. The vision (corrected if necessary) of each observer was checked, with three normalized tests:

1. Monoyer chart test was performed to check the acuity. All the subjects who took part in the experiment got a result higher than 9/10.
2. Ishihara test was performed to check color vision. Only the subjects without any color trouble took part in the experiment .
3. Randot stereo test was performed to check the 3D acuity. All the subjects who took part in the experiment got a result higher than 7/10.

6.3.2 Viewing conditions

Stimuli were displayed on a 26-inch (552×323 mm) Panasonic BT-3DL2550 stereoscopic LCD screen. Stereoscopy was achieved thanks to a pair of passive polarized glasses. The screen had a resolution of 1920×1200 pixels, and the refresh rate was 60 Hz. The maximum luminance of the display was 180 cd/m^2, which yielded a maximum luminance of about 60 cd/m^2 when watched through glasses. The environment luminance was adjusted according to each observer, in order to let the pupil have an appropriate size for eye-tracking. SMI RED 500 remote eye-tracker was used to record the eye movements. The accuracy of this eye-tracker is 0.4 degree.

The viewing distance has been set to 93 cm. In this condition, each pixel subtends about 62 arcsec and the whole screen subtends 33.06×18.92 visual degrees in the observer's field of view. All the objects were displayed in an area within 10.32×5.91 degrees. A chin-rest was used to stabilize observer's head, and the observers were instructed to "view anywhere on the screen as they want".

All 118 scenes were presented in random order. Each scene was presented for 3 seconds. After each scene, a point located in the center of the screen and with no disparity was presented for 500 ms. A nine-point calibration of the eye-tracker was performed at the beginning of the experiment, and repeated every twenty scenes. The quality of the calibration was verified by the experimenter on another monitor. Participants could allow themselves a rest before each calibration.

6.3.3 Stimuli

The experiment consisted of the presentation of stereoscopic scenes in which some identical objects were displayed at different depth planes.

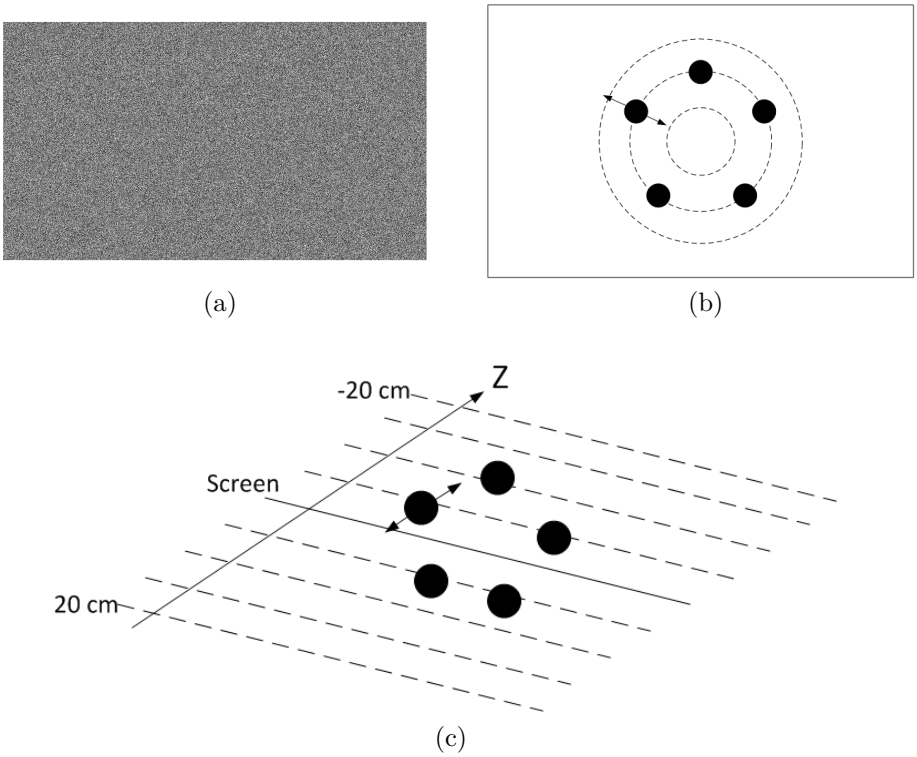


Figure 6.3.1: Composition of the stimuli. (a) The background of the stimuli. Only white-noise is contained in the background, which is positioned behind all the objects at -20 cm. (b) Positions of the objects' projections on the screen plane. Projected objects laid uniformly on a circle which centers at the screen center. (c) Allocation of objects in depth, potential depth values range from -20 cm to 20 cm by a step of 5 cm.

The background was a flat image consisting of white noise as shown in figure 6.3.1a, which was placed at a depth value of -20 cm (20 cm beyond the screen plane). In each scene, the objects consisted of a set of black disks of the same diameter S. They were displayed at different depth values randomly chosen among {-20, -15, -10, -5, 0, 5, 10, 15, 20} cm. Though the objects were placed at different depths (figure 6.3.1c), the positions of the objects' projection on the screen plane uniformly laid on a circle centered on the screen center (figure 6.3.1b). Thus, it can be assumed that no "center-bias" was introduced in the observation.

For the stereo viewing, the perceived depth is achieved by horizontally shifting the object towards different directions in the left view and the right view to simulate the binocular disparity. The relationship between disparity (in pixel) and the perceived depth (in cm) is modeled by Equation 6.3.1:

$$D = V/(1 + \frac{I \cdot R_x}{P \cdot W}) \tag{6.3.1}$$

where D represents the perceived depth, V represents the viewing distance between observer and screen plane, I represents the interocular distance, W and R_x represents, respectively, the width (in cm) and the horizontal resolution of the screen, P is the disparity in pixels. Note that a positive disparity value ($P > 0$) indicates a crossed disparity, and a negative disparity value ($P < 0$) indicates an uncrossed disparity. The objects at positive depth value are between the viewer and the display, while the objects at negative depth values are beyond the display. A depth value of 0 corresponds to the screen plane.

The depth range (from -20 cm to 20 cm) was chosen in order to match the comfortable viewing zone [Chen 10], in the particular viewing conditions of this experiment. Therefore, it could be assumed that the conflict between accommodation and vergence in our experiment would not cause unacceptable level of visual discomfort or visual fatigue [Hoffman 08].

To generate different stimuli, three parameters were independently varied from one scene to another:

- The number of objects, $N \in \{5, 6, 7, 8, 9\}$.
- The radius R of the circle on which the objects are projected on the screen plane, $R \in \{200, 250, 300\}$ pixels.
- The size of the objects, which is represented by the diameter of the disk S varying from 48 pixels to 168 pixels by a step of 12 pixels. Note that given a combination of N and S, the value of S was selected from the range of $[\frac{\pi R}{N\sqrt{2}}, \frac{2\pi R}{N\sqrt{2}}]$. The range serves to avoid any overlap between objects.

Derived from the combinations of this set of parameters, 118 scenes were presented to each observer. We had 30 five-object stimuli, 26 six-object stimuli, 23 seven-object

stimuli, 21 eight-object stimuli, and 18 nine-object stimuli. Figure 6.3.2 gives some examples of these different scenes. Note that the set of three independent parameters enabled the potential study of the impact of different factors on depth-bias. However, in the scope of this study, we particularly focus on the impact of objects number on depth-bias. We thus separated all the stimuli into five groups only depending on the number of objects (regardless the other two parameters).

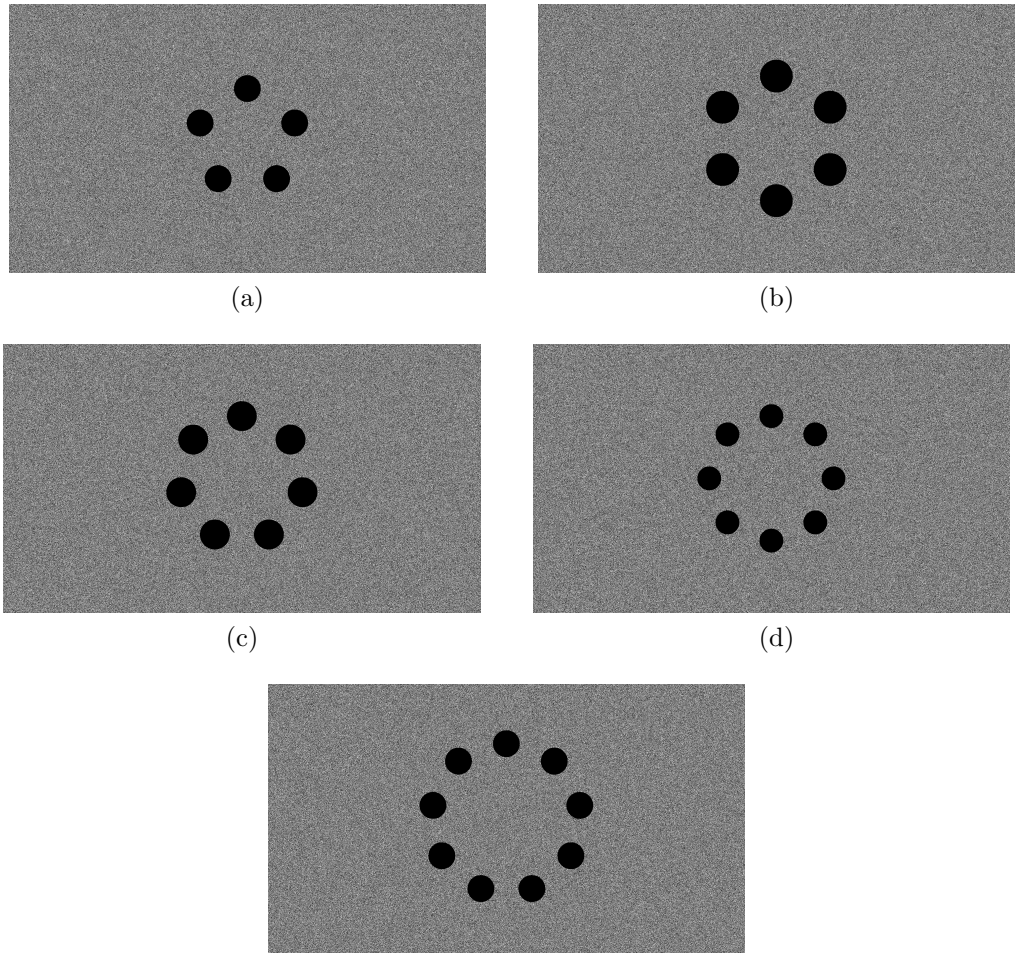


Figure 6.3.2: Examples of the five stimulus types (in terms of the number of objects contained) used in the eye-tracking experiment. These stimuli contain different numbers of objects.

There were several advantages of using this type of synthetic stereoscopic stimuli for the investigation of depth-bias:

- Firstly, it allowed a precise control of the influence of 2D visual features on visual attention. Even in 3D viewing, human's eye movements are still affected by many

bottom-up 2D visual features of the stimuli, such as color, intensity, object's size, and the center-bias. These factors could contaminate our evaluation of the influence of depth on visual attention. In our experiment, all the objects were with a constant shape, a constant size, and were positioned at a constant distance to the center of the screen. This set up let the stimuli get rid of as many bottom-up visual attention features as possible. The white noise background and the simple allocation of the objects allowed to avoid, as much as possible, the potential influence of top-down mechanisms of visual attention.

- Secondly, it allowed a precise control of the influence of depth cues on depth perception. Disparity was the only depth cue we took advantage of in this experiment. The reason of choosing binocular disparity was that the relationship between this depth cue and the perceived depth could be well modeled (see Equation 6.3.1). While for some other (monocular) depth cues, such as blur, perspective, occlusion [Wang 11b], the influence on perceived depth was more difficult to be quantitatively measured.

- Thirdly, the white noise background and the simple allocation of objects limited the complexity of the scenes presented to the observers to a low level, which made a shorter observation duration feasible. The viewing time in our experiment was short (3 seconds for each trail). Nevertheless, it was still long enough for participants to subconsciously position their fixations on objects and explore the scene as they want. Hence, using these simple stimuli allowed experimenters to collect more data, as well as to learn the evolution of depth-bias over time.

6.3.4 Post processing of eye tracking data

The first step of processing was to identify the fixations and filter out the saccades. The recorded eye movement data were processed by the event detection software "BeGazed" provided by SMI. This software selected saccades as primary events using a velocity-based algorithm [Salvucci 00]. According to this algorithm, fixations (and blinks) were computed and derived from the primary saccade events:

1. The velocities of all the recorded gaze points were first calculated. The peaks were then detected from all these velocities. Note that a "peak" was defined as the peak value of velocity above the Peak Threshold (i.e. 40 degree/s in our experiment). Given the stream of velocity values, for each peak, we searched to the left for the first velocity which was lower than the fixation velocity threshold, in order to detect the start of a saccade-like event. Similarly, we searched also to the right for the fist velocity lower than the fixation velocity threshold, in order to detect the end of the saccade-like event.

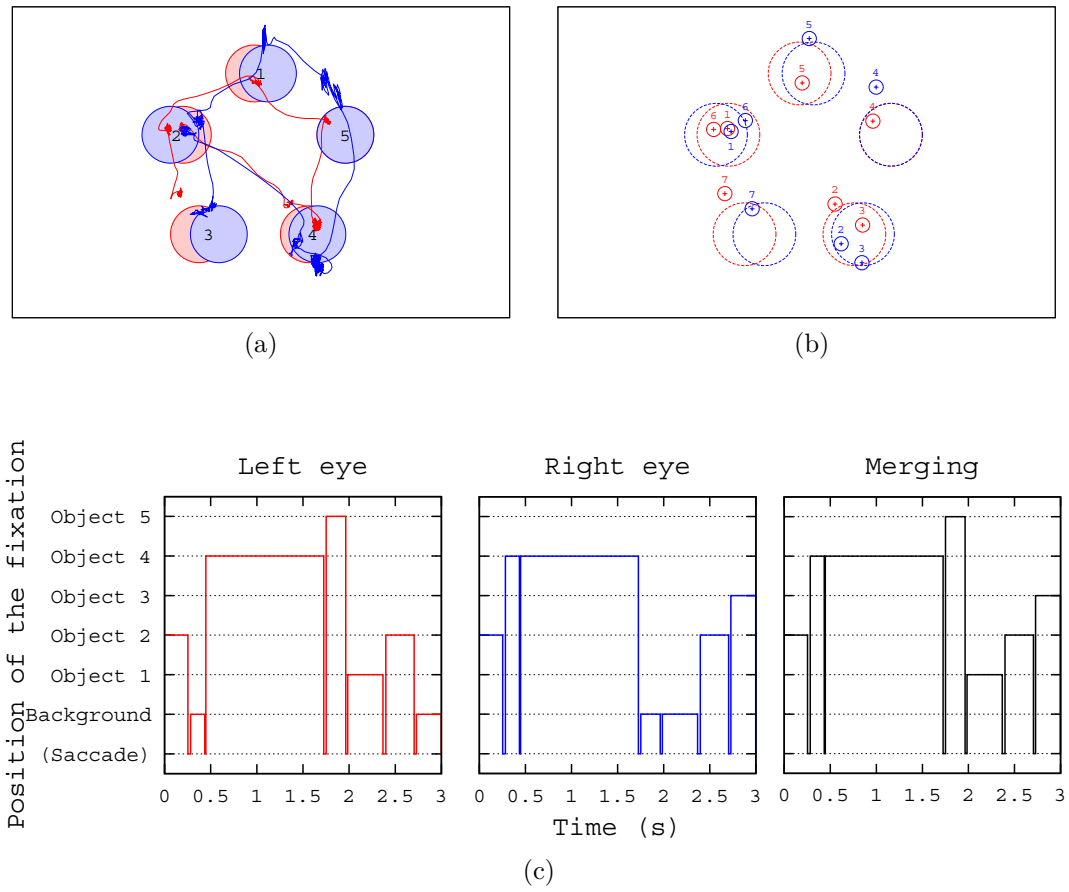


Figure 6.3.3: Illustration of the processes of relating a fixation to an object. In (a), the curves illustrate the path of each eye, and disks illustrate the objects in each view. Red color represents the left eye's path and the left view stimuli, while blue color represents the right eye's path and the right view stimuli. The numbers inside each object indicate objects' ID. (b) shows the fixations of each eye (same color as introduced previously) along with the area of detection for each object and each view. The detection areas are 10% larger than the size of the objects. The numbers indicate the temporal order of the fixations. (c) represents the position of each fixation as a function of time. Results are given for each eye, and for the merging of both eyes.

2. We assumed the saccade-like event a real saccade, if (1) the distance between start and end exceeded the Minimum Saccade Duration (22 ms), and (2) the single peak value lied in the range of 20% to 80% of the distance between start and end.

3. Finally, a fixation event was created between the newly and the previously created blink or saccade. All fixations below the minimum fixation duration (50 ms) were rejected.

The second step was to determine the spatial position of each fixation in order to relate it to the objects present in the scene. We found that directly computing the depth of a fixation based on the left and the right fixation's coordinates on the screen plane was difficult, due to the insufficient accuracy of the eye-tracker (see Figure 6.3.3a). Therefore, we adopted an indirect method to compute each fixation's depth. The computation was done independently for both eyes. Left eye positions were matched with left eye stimuli, and right eye positions were matched with right eye stimuli. It was then checked if a fixation was located on one of the objects or not. For each eye, a fixation was considered to be located on an object if it was positioned inside the object or in a surrounding area 10% larger than the object (to compensate for potential inaccuracy of the eye-tracker). Otherwise, the fixation was considered to be located on the background. Figure 6.3.3 illustrates the process.

Both eyes' fixations were then merged by the following rule: a given object was considered as being fixated if at least one eye's fixation was inside this object (Figure 6.3.3c). Because each object's depth is known, the depth of a fixation could be deducted from its position. Note that only the fixations located on a object were considered in the following analysis.

6.4 Results

6.4.1 Fixation distribution in depth

The number of fixations located on each object were calculated for each observer and each scene. The result was then transformed into a frequency distribution: for each observer, we divided the number of fixations on each object by the total number of fixations. We considered the uniform probability distribution $P_r = 1/N$, as a reference, based on the assumption that each object would attract the same amount of fixations if there was no depth-bias on the distribution of fixations (N represents the number of objects contained in a scene). This process was done repeatedly for all the five types of stimuli which contained different numbers of objects in the scene.

Figure 6.4.1 shows how the fixations are distributed on objects located at different depth planes in a scene. As we can see in the different plots, regardless the number of objects contained in the scene, the object closest to the observer always attracts most fixations (more than 20% of the total amount). The percentage of fixations then decreases

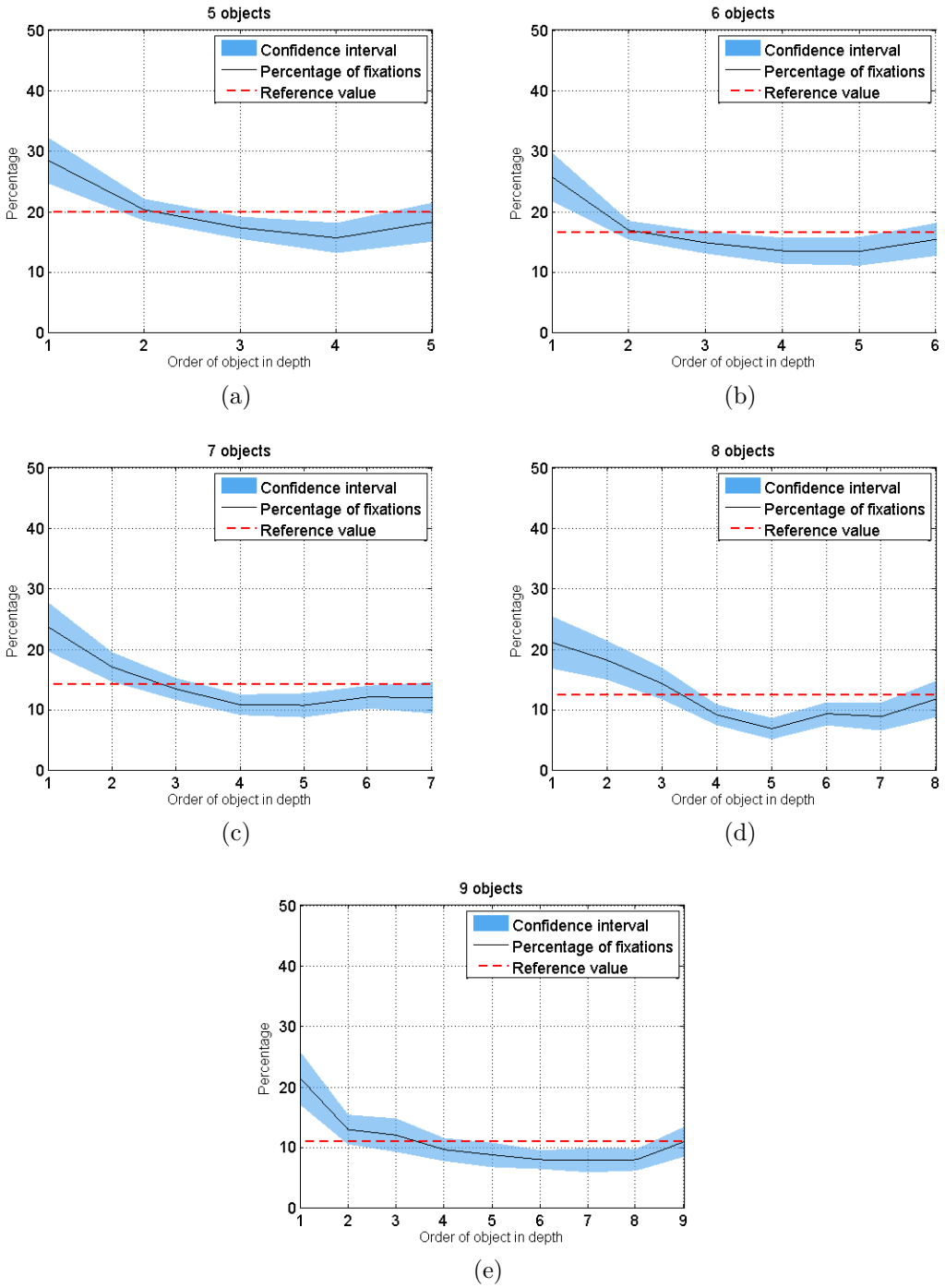


Figure 6.4.1: Fixation distribution (all the fixations were considered) as a function of the order of object's depth for the scenes containing different number of objects ($N \in 5, 6, 7, 8, 9$). X axis is the order of objects; sub-figure (a) to (e) repesents the group of scenes that contain 5 to 9 objects, respectively. Y axis represents the percentage of number of fixations. The blue area represents the 95% confidence interval. The dash line represents the uniform probability distribution (1/N).

as the depth order increases in the front half part of the scene. The curves generally follow a very similar shape in all five conditions. The frequency of the fixations on the objects located in the front range of the scene is significantly higher than predicted by the uniform probability distribution. The observation from these curves having similar trend supports the existence of depth bias.

Num. of objects	ANOVA result
5	$F(4, 130) = 11.73$, $p < 0.05$
6	$F(5, 156) = 12.42$, $p < 0.05$
7	$F(6, 182) = 13.22$, $p < 0.05$
8	$F(7, 208) = 13.80$, $p < 0.05$
9	$F(8, 234) = 11.85$, $p < 0.05$

Table 6.1: Results of the ANOVA performed on the fixation distributions presented in Figure 6.4.1.

A one-way ANOVA has been performed to check the statistical significance of the values for each type of the scene. The results (presented in Table 6.1) confirm that there exists a significant effect of fixation's depth order on fixation distribution. A post hoc paired t-test with Bonferroni correction has been then performed to check the significant difference between each pair of ordinal fixations in depth. For all the conditions, the percentage of fixations on the first object is significantly higher than the others, while the fixation percentage from the third to N^{th} ordinal objects are not significantly different from each other.

6.4.2 Variation of fixation's depth as the function of fixation's temporal order

The curves in Figure 6.4.2 show how the first fixation of all observers is distributed on the objects in each type of stimuli. These curves were computed in the same way as introduced in the previous section, except that only the first fixation of each observation was considered. These curves indicate the degree of depth-bias in a short viewing duration at the very beginning of observation. If we compare each curve in Figure 6.4.2 to the corresponding one in Figure 6.4.1, we find that the first fixations are more likely located on the closest object in a scene. These distributions of the first fixations demonstrate that the first fixation on each stimuli is with a higher-degree of depth-bias than the following fixations. This observation demonstrates the evolution of depth-bias as the temporal order of fixation increases.

To further evaluate the temporal evolution of the fixations' average depth, we investigated how the average depth of fixations varied as the temporal order of fixation increases. The relative depth position of each fixation in the scene's depth range was first computed by equation 6.4.1

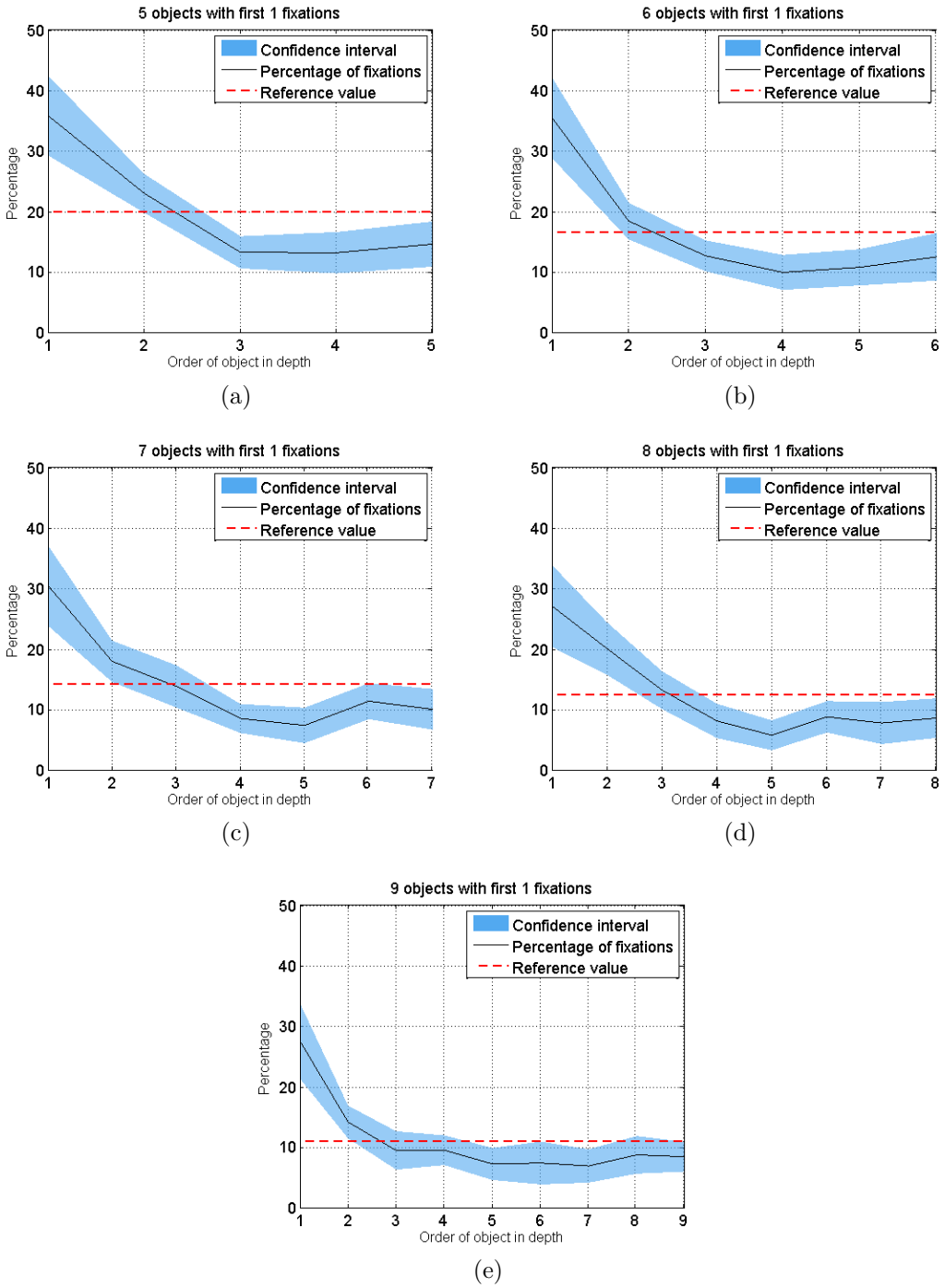


Figure 6.4.2: Fixation distribution (only the first fixation was considered) as a function of the order of object's depth for the scenes containing different numbers of objects ($N \in 5, 6, 7, 8, 9$). X axis is the order of objects; sub-figure (a) to (e) represent the group of scenes that contain 5 to 9 objects, respectively. Y axis represents the percentage of number of fixations. The blue area represents the 95% confidence interval. The dash line represents the uniform probability distribution ($1/N$).

$$D_{r_i} = (D_i - D_{min})/(D_{max} - D_{min}) \tag{6.4.1}$$

where D_i is the absolute depth of the i^{th} fixation, D_{min} and D_{max} are the minimum and maximum absolute depth of objects in a scene, respectively. Relative depth of the first seven fixations that were located on objects are computed and plotted in Figure 6.4.3.

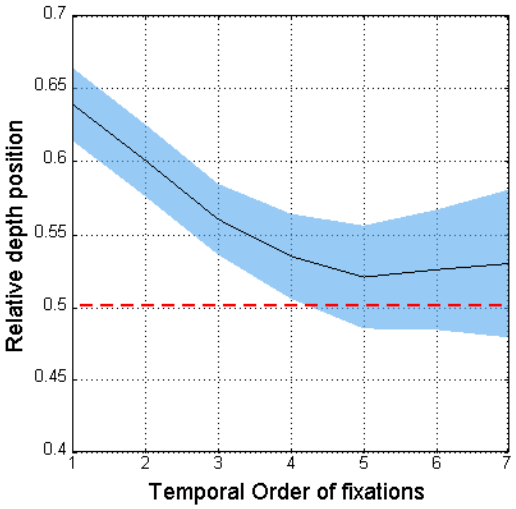


Figure 6.4.3: Relative depth position of fixations as a function of temporal order. The red dash line indicates the average value of relative depth of all objects displayed in the experiment. The light blue region represents the 95% confident interval.

An initial front response upon a new scene is observed in every participant and revealed in Figure 6.4.3. In this figure, the red dash line is plotted to indicate the average value of relative depth of all objects displayed in the experiment. If there was no depth-bias, the observers would explore the scene uniformly in depth during the observation, and each object in the scene should have the same probability to be fixated. That means the average depth value should vary little throughout the fixation sequence, and stay around the value of 0.5. However, a clear decrease of the average depth can be observed in figure 6.4.3. The first fixations are also found to be more often located on the objects close to the viewer. A one-way ANOVA is performed to check the significant difference among the relative depth of each temporally ordinal fixation. The result shows an effect of fixation sequence on depth ($F(6,860) = 7.94$, $p < 0.01$). A post hoc paired t-test with Bonferroni correction shows that the depth of the first and second fixation is significantly higher than the following fixations ($p < 0.01$).

This curve of fixations' average depth as a function of fixations' temporal order shows a viewing strategy that observers tend to explore a scene from the closest objects. The average depth values of all fixations are higher than the average depth of objects, which

means that observers pay more attention to the objects in the front part of the scene than the objects in the back part. All these observations support the existence of depth-bias.

6.4.3 Time dependence of fixation distribution in depth

The analyses in the previous sections reveal a variation of fixations' average depth according to the temporal ordinal of fixations. This variation implies that the level of depth-bias may be time dependent. In order to verify the time-dependence, we uniformly separated the 3-second observation time into six slices, then the fixation distribution as a function of depth (as introduced in Section 6.4.1) was computed for each slice of time. This processing was done repeatedly for all the five types of stimuli. The results are illustrated in Figure 6.4.4.

As demonstrated in Figure 6.4.4, in all the five types of scene, a clear depth-bias is found within the first 1000 ms observation time.

As the observation time increases, the number of fixations located on the closest object becomes smaller. The distribution of fixations on all the objects in a scene becomes more uniform. This tendency holds for all the five types of stimuli regardless the numbers of objects contained in the scenes. However, even if it is clear that the depth bias occurs at the beginning of the presentation, it is still hard to draw a conclusion of the time-dependence of depth-bias. Since once an object has been looked at, it is less likely to be looked at again due to the inhibition of return [Klein 00].

6.4.4 Latencies of fixations on objects at different depth

The latencies that each object got fixated for the first time are plotted as a function of object's depth order. Figure 6.4.5 illustrates separately the results for the scenes containing different numbers of objects. These plots visualize the tendency that the fixations hit the close objects earlier than the distant objects. The latency that all the objects got fixated ranged from 900 ms to 1800 ms. Generally, the object closest to the observer was fixated with the shortest latency (around 1000 ms). As the depth of object increased, the latencies of fixations hitting these objects varied positively.

6.4.5 Variation of depth-bias among individuals

The previous analyses demonstrate the existence of depth-bias, especially at the very beginning of the observation. The close objects generally attracted more and earlier fixations than the distant objects. However, inter-observer variations in fixation distribution in depth was also observed in our experiment. By analyzing how each observer's first fixation distributes, we found out that some observers did not prefer to start the observation from the close objects.

For each observation, the relative depth of each observer's first fixation was first computed using the Equation 6.4.1. All these fixations were then classified into three

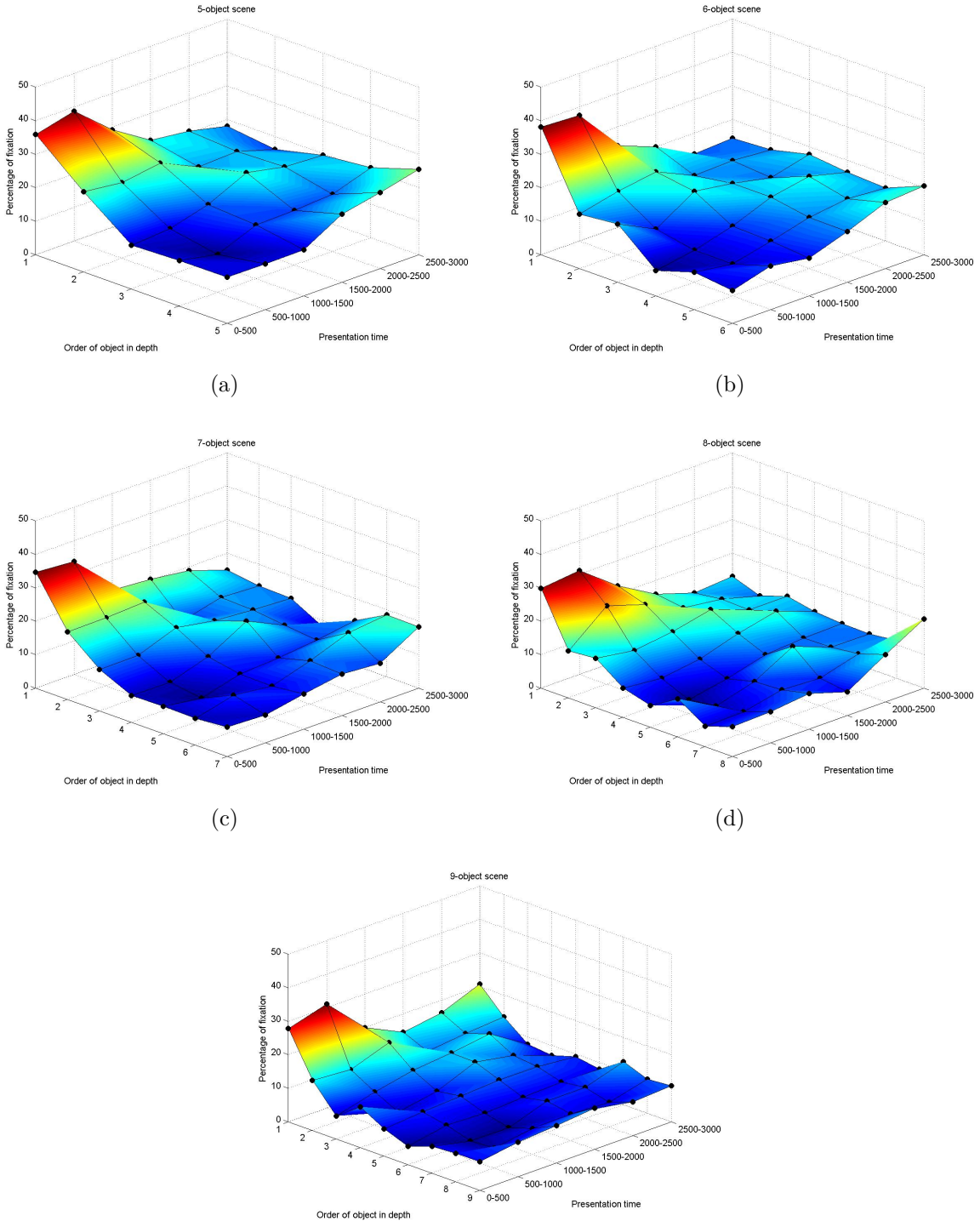


Figure 6.4.4: Variation of fixation distribution as a function of presentation time. The five sub figures represent five type of scenes which contain different numbers of objects.

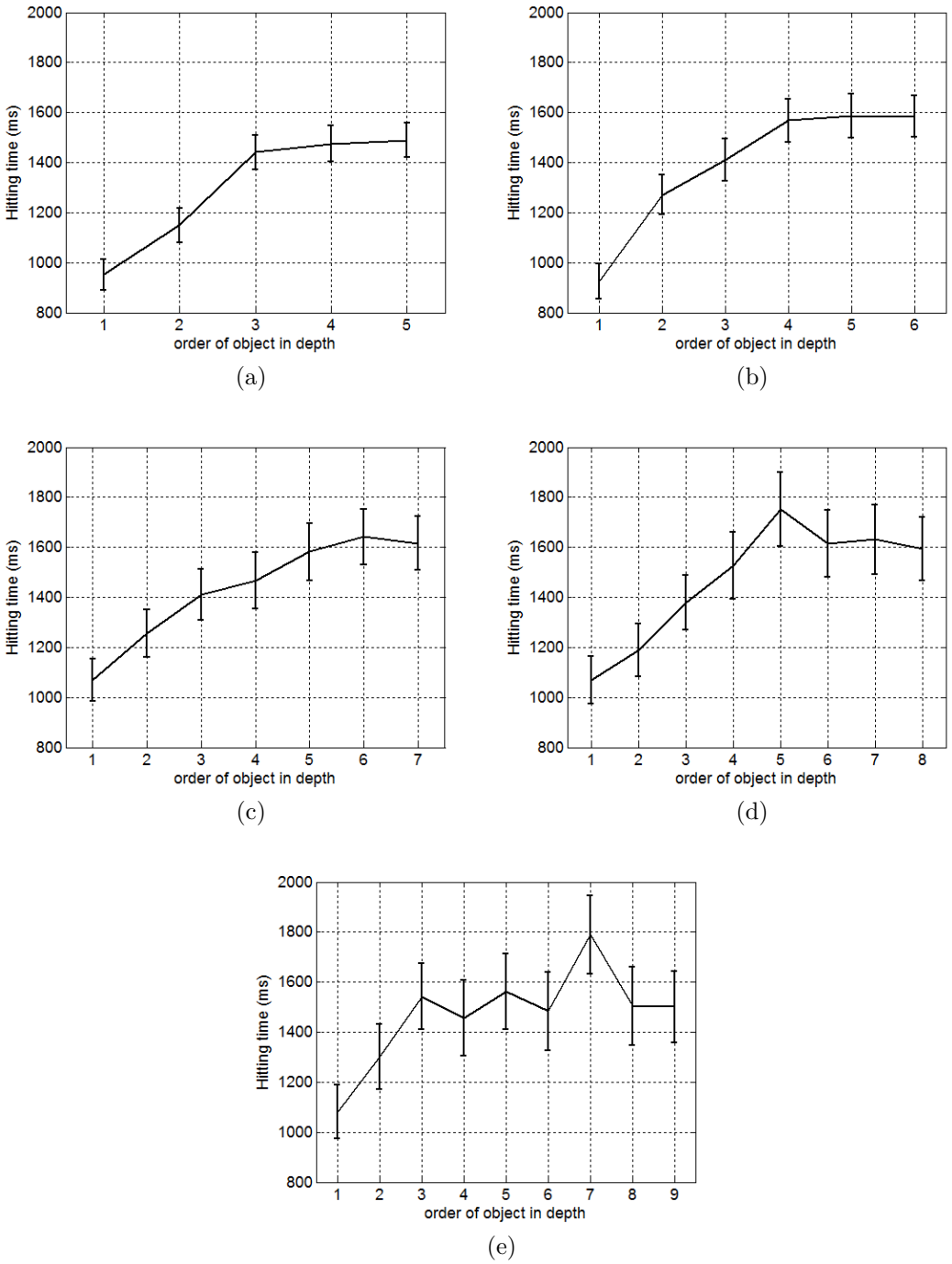


Figure 6.4.5: The latencies that each objects are fixated as a function of object's order in depth. The sub-figure (a) to (e) present the results for the scenes containing 5 to 9 objects, respectively. Y axis represents the latencies of the first fixation on a given object. X axis represents the order of objects. The error bar indicates the 95% confidence interval.

categories: the near fixations (having a relative depth ranging from 2/3 to 1); the middle fixation (having a relative depth ranging from 1/3 to 2/3); and the far fixations (having a relative depth ranging from 0 to 1/3). The percentage of each of the three types of fixations for each observer was then computed.

This classification resulted in an entity P containing three elements per observer:

$$P(observer) = [p_{near}, p_{middle}, p_{far}] \tag{6.4.2}$$

where p_{near}, p_{middle}, p_{far} represent the percentage of fixations (per observer) belonging to "near fixation", "middle fixation", and "far fixation", respectively. Figure 6.4.6a illustrates the three-dimensional scatter plot of the entity P for the 27 observers. This plot visualizes observer clustering patterns. Based on correlation distance, we used K-means to classify the observers into three groups. For each group of observers, a paired-sample t-test with Bonferroni correction was used to test the statistical significance for the difference among the percentages of different types of fixations:

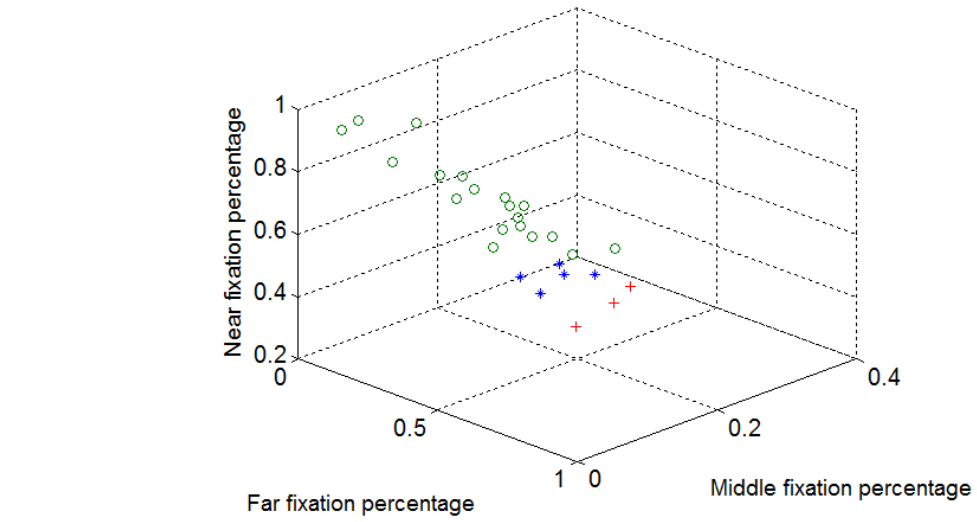
- Group 1 contains 19 observers. This group of observers had a strong depth-bias to the objects in the front part of the scene (see Figure 6.4.6b). The percentage of near fixations was significantly higher than the percentage of the others two types of fixations. There was no significant difference between the percentages of far fixations and middle fixations.

- Group 2 contains 5 observers. These observers had a depth-bias to some closest objects and some most distant objects (see Figure 6.4.6c). The percentage of far fixations and near fixations was significant higher than the middle fixations. However, the objects in the middle part attracted less fixation from this group of observers. There was no significant difference between the far fixations and near fixations.

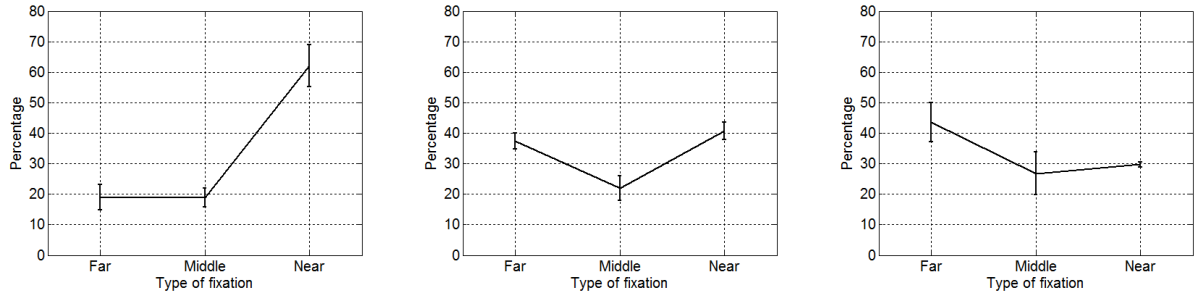
- Group 3 contains 3 observers. These three observers had an obvious depth-bias to the distant objects (see Figure 6.4.6d). The percentage of distant fixations was significantly higher than the percentage of the others two types of fixations. There was no significant difference between the percentages of near fixations and middle fixations.

6.5 Discussion

The main goal of the present study is to determine if there exists a so-called "depth-bias" in the viewing of 3D content on planar stereoscopic display. We examined how the depth order and the relative depth of the objects influenced observers' viewing behavior. Experimental results clearly show that observers payed more and earlier attention to the objects closest to them than to the other objects. This phenomenon could be



(a)



(b) (c) (d)

Figure 6.4.6: (a) The three-dimensional scatter plot of the percentage of three categories of fixations obtained from 27 observers. (b), (c) and (d) shows the percentage of each type of fixations obtained from the three groups of observers.

caused by a viewing strategy that people prefer to explore a scene from the objects with least distance, since this kind of objects may mean some potential dangers in the nature [Bowler 89]. We found that the depth-bias was obvious at the very beginning of observation(i.e. the first fixation of each observation). This initial bias and the short reaction time implies that depth-bias might be the result of a bottom-up mechanism.

Results also showed that the preference of looking at the closest objects decreased as the observation time increased. After one second of observation, fixations were distributed almost equally on all the objects in the scene regardless of the depth order of these objects. However, the sparse nature of stimuli makes it hard to draw conclusions on the time-dependence of depth-bias. In the experiment, once the closest object had been looked at, other objects would be looked at and it appeared that they were selected in no particular depth order. This variation of fixations' average depth could be caused by the inhibition of return [Klein 00].

On the other hand, we found that the effect of depth-bias was not consistent for all observers. To most of the observers, the object closest to them attracted most fixations. However, for some other observers, the distant objects (or both the closest object and the distant objects) attracted more fixations than the objects in the middle part. This inconsistency of preference implied the different viewing strategies in observer in the viewing of stereoscopic 3D scene.

In the present study, the synthetic stimuli used were designed to get rid of (as much as possible) the effect of the visual features other than depth. However, this is not the only approach. Despite the advantages of using the proposed synthetic stimuli, there exists another approach to measure the effect of depth: using stimuli taking into account also the other sources of visual information. This alternative approach can demonstrate the contribution of depth above and beyond that of other factors. The present study shows that the depth-bias seems to be a bottom-up process. Therefore, this alternative approach might help to verify if the effect of depth-bias is still significant in presence of top-down information (i.e. the semantics of picture).

In terms of applications, studying the depth-bias can be beneficial to the development of computational models of 3D visual attention for both still images and videos. In the area of computationally modeling of 3D visual attention, the depth-bias is usually considered to be linear to the disparity and time-independent [Maki 96b, Zhang 10, Chamaret 10]. Our results about the fixation distribution in depth and the time-dependence of the depth-bias might be used to improve these computational models in predicting salient areas of a 3D scene.

In the literature, studies have shown that stereoscopic vision relies mainly on relative depth difference between objects rather than on their absolute distance in depth from where the eyes fixate [Neri 04]. In the present study, the influence of depth on the distribution of visual attention is examined based on relative depth. However, the study of the influence of absolute depth information is not included in the scope of this chapter. Absolute depth, which is linear to the binocular disparity provided by stereoscopic displays, can be another important factor affecting the depth-bias, since it has

been demonstrated the existence of disparity-selective neurons in the primary visual cortex (V1) [Neri 99, Barlow 67, Nikara 68, Poggio 77]. Moreover, we also know that the conflict between accommodation and vergence, which is caused by binocular disparity, affects the viewing behavior of stereoscopic content. The existence of areas with disparity larger than a threshold can cause the problem of visual fatigue and visual discomfort. Obviously, this type of area will not attract too much attention, even if it is salient in terms of visual feature. All these evidences testify to the influence of the absolute depth (i.e. disparity) on visual attention in the viewing of stereoscopic 3D stimuli.

6.6 Conclusion

In this study, we conducted an eye-tracking experiment using state-of-the-art stereoscopic display and eye-tracker. A large number of synthetic stimuli were designed for the experiment in order to get rid of the effect of 2D visual features, and let the visual attention of observers be influenced by only depth information. Results demonstrate the existence of the depth-bias in the viewing of 3D content on a planar stereoscopic screen.

Key points

Context

- ❏ Observers' fixations exhibit a marked bias towards certain areas on the screen. The center-bias is an instance of this phenomenon. Moreover, in the viewing of 3D content, a bias regarding to the depth information has been found. This bias is named as "depth-bias".
- ❏ Depth-bias has been mentioned and applied in previous studies of computational modeling of 3D visual attention. However, the hypothesis that depth-bias is linear is still arbitrary.
- ❏ In order to study the effect of depth-bias, stimuli used in the eye-tracking experiment need to be well designed, in order to get rid of the effect of the visual features other than depth information. However, this kind of psychophysical study is still lacking.

Contributions

- ❏ We conduct an eye-tracking experiment using synthetic stimuli, state-of-the-art stereoscopic display and eye tracking system.
- ❏ We design a large number of synthetic stimuli for the eye-tracking experiment. In these stimuli, the effect of 2D visual features is largely avoided. In the viewing of these stimuli, observers' visual attention is mainly influenced by only depth information.
- ❏ Our results demonstrate the existence of the depth-bias in the viewing of 3D content on a planar stereoscopic screen. This depth-bias is shown to be time-dependent. We also investigate the variation of depth-bias among individuals.

Chapter 7

Computational modeling of stereoscopic 3D visual attention

A key point of the studies regarding computational models of 3D visual attention is the development of efficient models that can well predict the distribution of saliency. After introducing the studies regarding (1) the ground truth of 3D visual attention and (2) the impact of depth perception on visual attention, we particularly focus on the modeling of 3D visual attention. A new 3D visual attention model relying on both 2D visual features and features extracted from depth information is proposed and evaluated in this chapter.

We first introduce and summarize previous works on 3D visual attention (Section 7.2 for psychophysical studies, and Section 7.3 for computational modeling). A taxonomy of computational models of 3D visual attention is proposed. In Section 7.4, we introduce a depth-saliency-based model of 3D visual attention. To benefit from psychophysical studies, we propose to apply Bayes's theorem on the result of an eye-tracking experiment using synthetic stimuli to model the correlation between depth features and the level of depth saliency. We also introduce and discuss two ways of combining depth saliency map with 2D saliency map. At the end of the chapter, conclusion and discussion are presented in Section 7.6.

7.1 Introduction

During viewing stereoscopic 3D content, disparity information is exploited by brain to retrieve the 3D layout of the environment and leads to a stereoscopic perception of depth. This change of depth perception also largely changes deployment of human's visual attention when watching stereoscopic 3D images/videos [Hakkinen 10, Huynh-Thu 11b]. Predicting the salient areas of a 3D scene becomes thus a challenging task due to the additional depth information.

Several challenges, importances and new applications of visual attention for 3D content viewing were introduced by Huynh-Thu et al. in [Huynh-Thu 11a]. They described the

conflicts that the human visual system has to deal with during watching 3D-TV, as well as how these conflicts might be limited and visual comfort could be improved by knowing how visual attention is deployed. Several new application areas that can be beneficial by being provided the location (including depth) of salient areas were also introduced. These candidate applications exist in the different steps of a typical 3D-TV delivery chain, e.g. 3D video capture, 2D to 3D conversion, reframing and depth adaptation, and subtitling in 3D movie.

The increasing demand of visual-attention-based applications for 3D content arises the importance of computationally modeling 3D visual attention. However, two questions need to be figured out for developing a 3D visual attention model:

- Influence of 2D visual features (e.g. color, intensity, orientation, center-bias) in 3D viewing condition.
- Influence of depth on visual attention deployment in 3D viewing condition. For instance, it is necessary to figure out how the bias of fixations according to depth (i.e. the depth-bias), and the visual features based on depth information (e.g., the orientation of surface, the contrast of depth) affect the deployment of human's visual attention.

In the last decade, a large number of 2D visual attention models have been investigated. Therefore, the first question concerns the possibility of adapting these large amount of existing 2D models into 3D case. On the other hand, the second question concerns the means by which the additional information, depth, can be taken into account.

The research of 3D visual attention modeling is also facing another problem: any published eye-tracking database of 3D still images is still lacking. In addition to the lack of quantitative evaluation of performance, another consequence of the lack of ground truth is that most of the existing 3D visual attention models only take into account, in a qualitative way, the results of psychophysical experiments about depth's influence or the variation of 2D features' effects. Any model that quantitatively integrates experimental observation results is still missing so far. Moreover, there is still not a strong conclusion on the means by which depth information should be used in 3D visual attention modeling: depth should be used to weight 2D saliency map; or alternatively be considered as an additional visual dimension to extract depth features and create depth saliency map.

7.2 How the deployment of 3D visual attention is affected by various visual features: previous experimental studies

Based on observations from psychophysical experiments, several studies have started to examine both qualitatively and quantitatively how visual attention may be influenced by the 2D visual features and additional binocular depth cue.

One of the early works was done by Jansen et al. [Jansen 09] who investigated the influence of disparity on viewing behavior in the observation of 2D and 3D still images. They conducted a free-viewing task on the 2D and 3D version of the same set of images. They found that the additional depth information led to an increased number of fixations, shorter and faster saccades, and increased spatial extent of exploration. However, no significant difference was found between the viewing of 2D and 3D stimuli in terms of saliency of several 2D visual features including mean luminance, luminance contrast, and texture contrast. This consistence of the influence of 2D low-level visual features implied: (1) the importance of 2D visual feature detection in the design of 3D visual attention model, and (2) the possibility of adapting existing 2D visual attention models in modeling of 3D visual attention.

Liu et al. [Liu 10] examined visual features at fixated positions for stereo images with natural content. Instead of comparing viewing behavior between 2D and 3D content viewing, they focused on comparing visual features extracted from fixations and random locations in the viewing of 3D still images. On one hand, they demonstrated that some 2D visual features including luminance contrast and luminance gradient were generally higher at fixated areas. On the other hand, their results also indicate that disparity contrast and disparity gradient of fixated locations are lower than randomly selected locations. This result is inconsistent with the result from Jansen et al. who found that observers consistently look more at depth discontinuities (high disparity contrast areas) than at planar surfaces. One limitation of Liu et al.'s study might rely on the quality of ground truth disparity map. The disparity maps they used came from a simple correspondence algorithm rather than any depth range sensing systems or any sophisticated depth estimation algorithms. The final results might thus be affected by a considerable amount of noise in the estimated disparity maps.

Hakkinen et al. [Hakkinen 10] examined the difference in eye movement patterns between the viewing of 2D and 3D versions of the same video content. They found that eye movements are more widely distributed for 3D content. Compared to the viewing of 2D content, viewers did not only look at the main actors but also looked at some other targets on typical movie content. Their result shows that depth information from the binocular depth cue provides viewers additional information, and thus creates new salient areas in a scene. This result suggests the existence of a saliency map from depth, and a potential "summation" operation during the integration of 2D and depth saliency information. In opposite, Ramasamy et al.'s study [Ramasamy 09], which is related to stereo-filmmaking, showed that observers' gaze points could be more concentrated when viewing the 3D version of some content (e.g. the scenes containing long deep hallway).

In terms of the depth plane where fixations tend to be located, Wang et al. [Wang 11b] examined a so-called 'depth-bias' in task-free viewing of still stereoscopic synthetic stimuli. They found that objects closest to the observer always attract most fixations. The number of fixations on each object decreases as the depth order of the object increases, except that the furthest object receives a little more fixations than the one or two objects in front of it. The number of fixations on objects at different depth planes were

also found to be time dependent. This result is consistent with the result of Jansen et al. [Jansen 09]. Considering the influence of center-bias in 2D visual attention, these results indicate the existence of a location prior according to depth in the viewing of 3D content. This location prior indicates the possibility of integrating depth information by means of doing a weighting.

Wismeijer et al. [Wismeijer 10] examined if saccades were aligned with individual depth cues or with a combination of depth cues by presenting stimuli in which monocular perspective cues and binocular disparity cues conflicted. Their results indicate a weighted linear combination of cues when the conflicts are small, and a cue dominance when the conflicts are large. They also found that vergence is dominated only by binocular disparity. Their result implies that the interocular distance recorded by binocular eye-tracking experiment for 3D content should be compensated by taking into account the local disparity value.

7.3 Previous works on 3D visual attention modeling

As introduced previously, great efforts have been put into the study of viewing behavior of 3D content. However, in terms of the development of computational models, only a few computational models of 3D visual attention have been proposed, compared to the body of 2D visual attention models. Experimental results have demonstrated strong influences of 2D visual features, in the viewing of 3D content. However, due to the addition of new depth cues, depth features, and their combination or conflicts [Hoffman 08, Okada 06] with other monocular cues, a direct use of 2D visual attention model for 3D content is neither biologically plausible nor effective.

Furthermore, the disparity between two views can raise serious challenges on collecting 3D gaze points and creating fixation density maps which are used as ground-truth, since the gaze data need to be extrapolated or processed to provide a notion of depth in relation with gaze direction or location [Huynh-Thu 11a].

In the literature, several computational models of 3D visual attention have been investigated. All of these models contain a stage in which 2D visual features are extracted and used to compute 2D saliency maps. However, according to the ways they use depth information, these models can be classified into three different categories: depth-weighting model, depth-saliency model, and stereo-vision model.

7.3.1 Depth-weighting models

This type of models (e.g. [Maki 96b], [Zhang 10] and [Chamaret 10]) do not contain any depth-map-based feature-extraction processes. Apart from detecting salient areas by 2D visual features, these models share a same step in which depth information is used as the weighting factor of the 2D saliency. The saliency of each location (e.g. pixel, target or depth plane) in the scene is directly related to its depth. Both 2D scene and

depth map are taken as input. Note that depth maps used in these models can be a ground truth depth map provided by depth detection equipments, or come from depth estimation algorithms which use two or multiple views.

Maki's model

One early computational model of depth-based visual attention is proposed by Maki et al.[Maki 96b]. The architecture of this model consists of a first stage of parallel detection of several preattentive cues (i.e. image flow, stereo disparity, and motion detection), followed by a stage in which different cues are integrated. A schematic diagram of the model is shown in Figure 7.3.1.

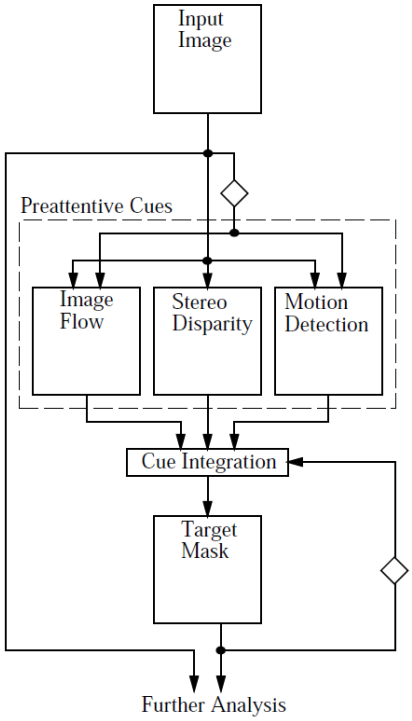


Figure 7.3.1: A schematic diagram of Maki et al.'s model.

In the early parallel stage, relative depth which plays a central role in the model is derived from a dense disparity map. The disparity map used in this model is derived from a phase-based algorithm [Maki. 96a], which is considered to have the advantages of low computational cost, stability against varying lighting conditions and especially allowing good direct localization of the estimated disparity.

The goal of the second stage, an integration stage, is to guide the attention to an appropriate part of the image given the information of different cues from the previous stage. The combination of different early cues is achieved in two modes, namely the pursuit and saccade modes, which are found in human scan paths. Each of these two

independent modes produces a target mask. The integration of cues uses selection criteria based on nearness and motion. Here, depth is used to apply a priority criterion based on the hypothesis that the target closer to the observer in depth is selected with higher priority. As discussed by the authors, this hypothesis can be reasonable in a scenario where the observer has to avoid obstacles.

However, the hypothesis made by Maki et al. may not necessarily hold in some conditions, for instance a scenario of viewing complex entertainment video content where the closest object may not be the only or main object of interest [Huynh-Thu 11a]. Actually, the authors showed that their model kept focusing on the moving object closest to the camera.

Zhang's model

Another proposal of bottom-up visual attention for 3D content comes from Zhang et al. [Zhang 10]. This model consists in extending a hierarchical model for 3D content by taking into account depth map as an addition cue (see Figure 7.3.2).

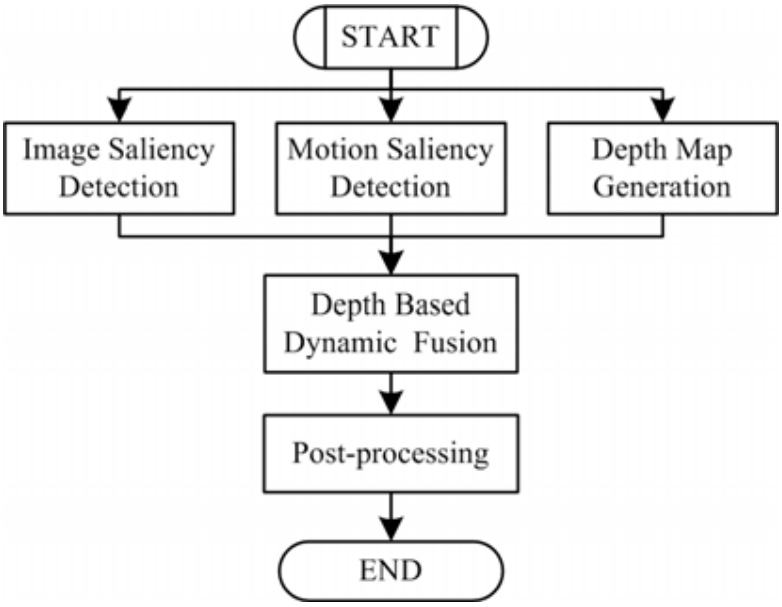


Figure 7.3.2: A schematic diagram of Zhang et al.'s model.

The 2D saliency detection of this model is achieved by adopting Itti's bottom-up attention model, in which color, intensity and orientation are taken as input. A motion saliency map based on motion contrast is also constructed.

In terms of the depth, the general hypothesis made by Zhang et al. is also that the level of interest of objects decreases as depth increases. The model firstly estimates the

disparity map for stereoscopic content, then converts the disparity of each pixel into perceive depth. The pixels with a small depth get a higher saliency. A depth-based fusion of 2D spatial saliency map and motion saliency map is then performed, based on the rule that the attractiveness of the stereoscopic objects decrease as they are getting further while within depth of field of the camera. Note that although depth is taken into account as an additional cue in Zhang et al.'s model, this model is still considered to be a depth-weighting model due to the hypothesis made by the authors, and the absence of a depth-feature extraction stage.

Chamaret's model

Chamaret et al. propose an adaptive 3D rendering method which is based on salient regions and thus contains a computational model of visual attention for stereoscopic 3D content [Chamaret 10]. To detect the salient regions of a 3D scene, the model described in [Le Meur 06] and [Le Meur 07] is first used to create a 2D saliency map based on low-level visual features. Note that, as discussed by the authors, this 2D saliency detection stage can adopt any 2D visual attention model including bottom-up models and top-down models. In terms of the way of integrating depth information, Chamaret et al. multiply the depth map with the 2D saliency map to maintain the common important pixels in the final map. Therefore, the diagram of the 3D visual attention model in Chamaret et al.'s adaptive 3D rendering method can be illustrated as in Figure 7.3.3.

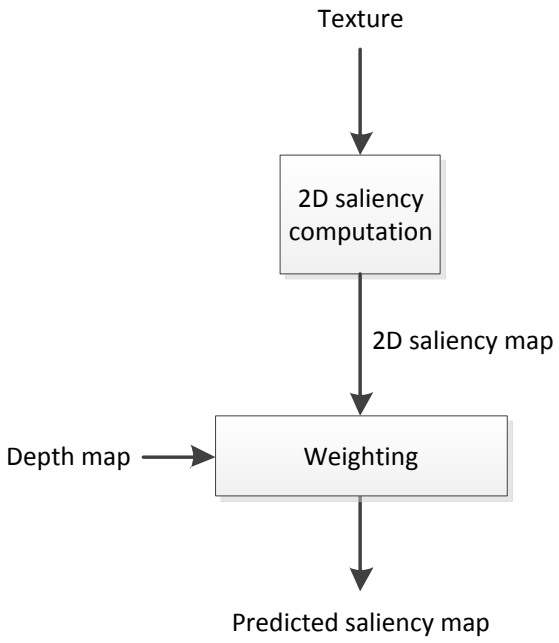


Figure 7.3.3: A schematic diagram of the 3D visual attention model in Chamaret et al.'s adaptive 3D rendering method.

7.3.2 Depth-saliency models.

Models (e.g. [Ouerhani 00] and [Potapova 11]) in this category take depth saliency as additional information. This type of models rely on the existence of "depth saliency maps". Depth features are first extracted from depth map to create additional feature maps, which are used to generate the depth saliency maps. These depth saliency maps are finally combined with 2D saliency maps (e.g. from 2D visual attention models using color, orientation or intensity) by using saliency map pooling strategy to obtain a final 3D saliency map. This type of model takes as input 2D scene and depth map.

Ouerhani's model

Ouerhani and Hugli proposed a computational model of 3D visual attention by using scene depth information to extend a bottom-up, task-independent, saliency-based 2D visual attention model [Ouerhani 00]. In terms of the 2D saliency detection, this model adopts the approach proposed by Itti et al. [Itti 98]. A number of low-level features are firstly extracted from the image to build feature maps. Based on a multi-resolution center-surround mechanism, these feature maps are then transformed into a set of corresponding conspicuity maps. The conspicuity maps are then pooled together in a competitive way. The diagram of Ouerhani and Hugli's model is illustrated in Figure 7.3.4.

In addition to the 2D visual features, additional depth-related features are extracted from the range image (i.e. depth map) of the scene. The authors supposed that the depth map is obtained by a 3D range camera. Three depth features were considered:

- Depth. This feature concerns the distance from camera to the objects in the scene, no additional operator is needed to extract this feature.
- Mean curvature. This feature concerns the surfaces in the scene. It is a second differential order feature providing information about the geometry of objects. However, the disadvantage of this feature is that it is sensitive to the noise in the depth map, and it is also sensitive to depth discontinuities. Therefore, smoothing operators need to be applied to the depth map to overcome the problem of noise. Depth discontinuities need also to be detected in order to compute this feature, mean curvature, only for continuous surfaces.
- Depth gradient. This feature is obtained by computing first order derivative on depth map. It can be an efficient way to detect depth changes in the scene.

However, in their paper, the usefulness of the depth features and the efficiency of their proposed model are only presented by showing a few images. There is no mention of any formal subjective experiment, such as an eye-tracking experiment, to evaluate the performance of the model and the added value of depth information.

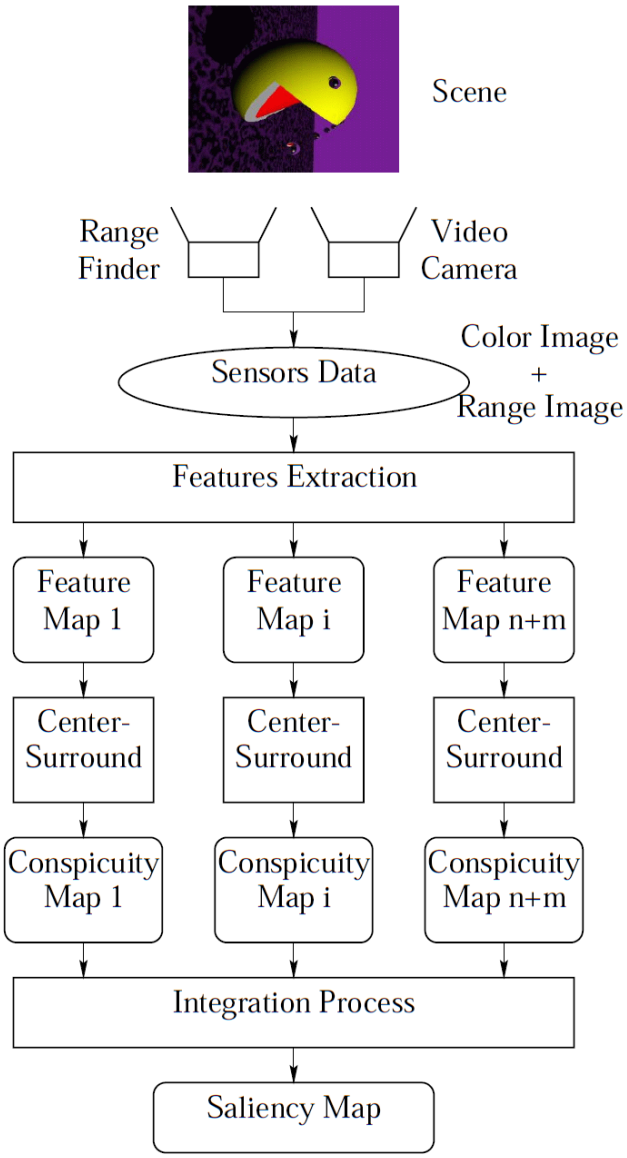


Figure 7.3.4: A schematic diagram of Ouerhani and Hugli's model.

Potapova's model

Potapova et al. propose a computational model of 3D visual attention based on probabilistic learning [Potapova 11]. The proposed model is designed for a robotic system. It takes into account not only the bottom-up attention and but also particular tasks, such as manipulation, grasping and tracking.

The authors use the Microsoft Kinect depth sensor to create an image database which contains 86 RGB images and the corresponding depth maps. Salient regions are hand-labeled by outlining them with a polygon. For each feature, the probability of observing that a pixel is marked as salient can be thus learned. Bayes rule is then used to infer the posterior probability, for each pixel, of being salient.

In Potapova et al.'s study, Itti's 2D bottom-up saliency model [Itti 98], which uses intensity, color and orientation as features, is deployed to predict 2D saliency map. Apart from the 2D visual features, several depth features are also investigated:

- Surface height. This feature is based on the hypothesis that (1) all the objects in the scene rest on a supporting plane, and (2) the objects sticking out of the clutter are more possible to be the objects of interest. To obtain this feature, the supporting plane, which is considered as the reference, have to be first detected. The distance of each point on the objects to the supporting plane are computed as the height. Each point is assigned a higher saliency value for a higher value of height.

- Relative surface orientation. This feature is based on the hypothesis that the objects' surfaces parallel to the supporting plane usually indicate top-surfaces of simple objects that are easier to be grasped. Relative orientation between the surfaces of objects and the supporting plane needs to be calculated to obtain this feature.

- Occluded edges. Canny operator is first applied on the depth map to detect the edges. Pixels (in the directions of decreasing depth value) closer to the edges are assigned higher saliency values.

Potapova et al. evaluate the performance of their data on the 3D images with hand-labeled salient regions. The results show that the two depth features, surface height and relative surface orientation, outperform simple 2D features. Occluded edges did not prove to be a useful cue for their task. They also investigate two ways to combine all the features, both of which have comparative performance.

Nevertheless, Potapova et al.'s model has some limitations which prevent it from being used as a general 3D visual attention model. In their study, the images used to train the model and assess the performance are limited to the scenes with some man-made objects on a table. Apart from that, their computation model is design for a robotic system to detect the objects that can be picked up in the scene, the depth features proposed in this work are search-task relevant features, instead of general bottom-up visual features.

Moreover, their model needs to be validated on the ground-truth data obtained from eye-tracker instead of hand-label experiment.

7.3.3 Stereo-vision models.

Instead of directly using depth map, this type of model takes into account the mechanisms of stereoscopic perception in human visual system. Bruce et Tsotsos propose a stereo-vision model of visual attention by extending an existing Selective Tuning model naturally to the binocular domain [Bruce 05]. They discussed the issue of binocular rivalry occurring in stereo-vision, as well as the difficult and biologically implausible translation of some types of 2D visual attention to the case of stereo-vision. They argued that a stereoscopic 3D visual attention model must take into account conflicts between the two eyes resulting from occlusions or large disparities.

Their model is developed based on the 2D model from [Tsotsos 95] using a visual pyramid processing architecture (see Figure 7.3.5). Images from both views are taken as input, from which 2D visual features can be considered. They add neuronal units into the visual pyramid processing architecture of the original 2D visual attention model for modeling stereo vision.

Bruce et Tsotsos demonstrate their model using a few simple synthetic images with structural objects composes of mainly straight lines. There is no mention of considerations about the images with more complex scenarios or natural content. On the other hand, they do not mention any comparison between their model's result with ground-truth data obtained from eye-tracking experiment.

7.3.4 Summary of the previous studies

Table 7.1 introduces main properties of the models belonging to each of the three categories. So far, most of the existing computational models of 3D visual attention belong to the first or the second category. Figure 7.3.6 summarizes the two different ways by which depth information is taken into account in these two types of models.

Both types of models have their respective advantages and limitations. Depth-weighting models can relatively easily adopt existing 2D models. The additional computational complexity is low due to the absence of depth feature extraction. However, a limitation of depth-weighting models is that they might fail in detecting some salient areas caused by depth features only. On the other hand, depth-saliency models use depth as an additional visual dimension, and take into account the influence of depth features by creating depth saliency maps. However, the consideration of depth features increases the computational complexity, and the influence of depth features on model's performance has not been quantitatively validated.

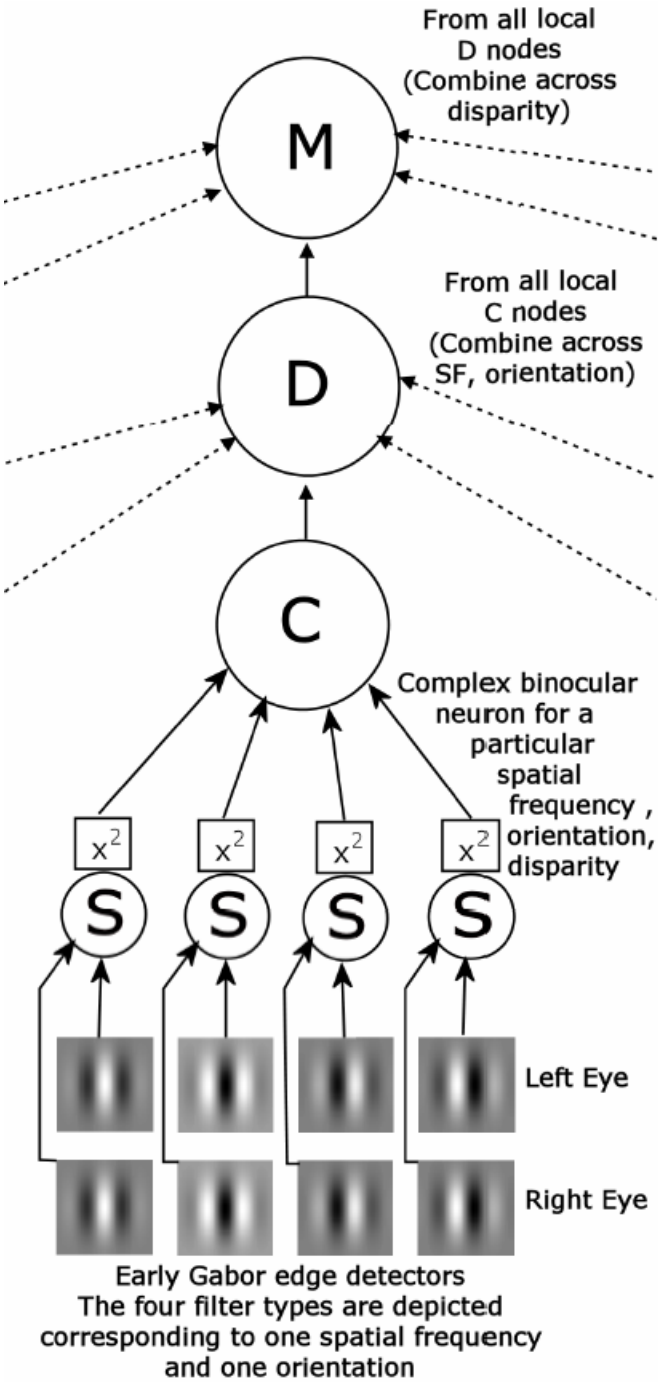


Figure 7.3.5: A schematic diagram of Bruce and Tsotsos's model.

DW	**Depth information**	**Operation**	**Validation**
Maki et al.	Relative depth	Assigned the target closer to observers with highest priority	Qualitative assessment; no quantitative comparison to eye-tracking data
Zhang et al.	Perceived depth, pop-out effect	Irregular space conversion. Pixels closer to observers and in front of the screen are considered to be higher salient	Qualitative assessment; no quantitative comparison to eye-tracking data
Chamaret et al.	Relative depth	Weight each pixel in 2D saliency map by its depth value	Qualitative assessment; no quantitative comparison to eye-tracking data
DF	**Depth information**	**Operation**	**Validation**
Ouerhani and Hugli	Absolute depth (distance), surface curvature, depth gradient	Extract depth features from depth map. Compute additional conspicuity maps based on depth features. Pool all the conspicuity maps (from 2D features and depth features)	Qualitative assessment; no quantitative comparison to eye-tracking data
Potapova et al.	Surface height, relative surface orientation, occluded edges	Compute one saliency map for each (2D and depth) feature, then sum all the saliency maps	Qualitative assessment and quantitative comparison to labeled ROIs
SV	**Depth information**	**Operation**	**Validation**
Bruce and Tsotsos	Disparity	Take two views as input. Add interpretive neuronal units for stereo-vision modeling into a 2D computational model which uses visual pyramid processing architecture.	Qualitative assessment; no quantitative comparison to eye-tracking data

Table 7.1: Main features of computational models of 3D visual attention. Note that DW denotes Depth-Weighting model, DF denotes Depth-Feature model, and SV denotes Stereo-Vision model.

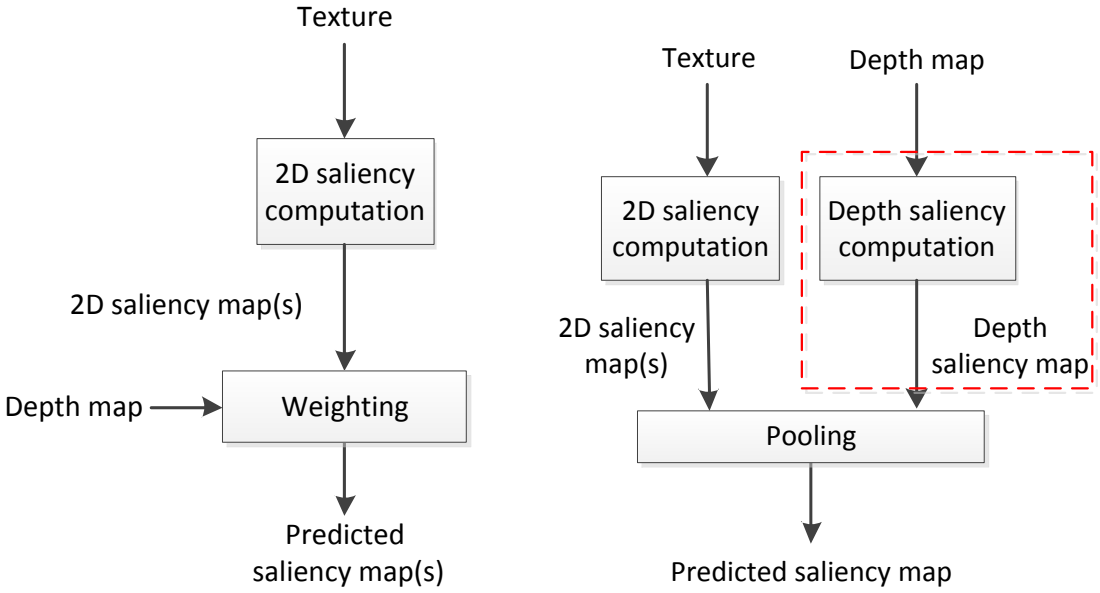


Figure 7.3.6: Two different ways of using depth in (a) the depth-weighting models and (b) the depth-saliency models. Note that the main difference between these two types of models is the existence of a stage for extracting depth features and creating depth saliency map.

7.4 A depth-saliency-based computational model of 3D visual attention

Depth features have been demonstrated their contribution on predicting saliency map of 3D images. Several depth features (e.g. surface curvature, depth gradient, relative surface orientation) have been proposed and used in previous 3D models. In this section, the way of creating depth map from which depth features can be extracted is first introduced. In the second step, we introduce a new method of generating a so-call "depth saliency map". The depth saliency map is computed based on only depth features (i.e. depth contrast) by using a Bayesian framework.

7.4.1 Depth map creation

We propose that a depth map providing the information of a scene's perceived depth needs to be computed at the first step of modeling 3D visual attention. In a stereoscopic 3D display system, depth information is usually represented by means of a disparity map which shows the parallax of each pixel between the left-view image and the right-view image. In literature, disparity map is usually directly adopted as depth information [Chamaret 10]. However, we propose that a transformation from disparity map to depth map which represents perceived depth in unit of length should be added in the chain of 3D visual attention modeling, since even the same disparity value corresponds to different perceived depth depending on the viewing condition.

From the view of display system, disparity is measured in unit of pixels. The relationship between disparity (in pixel) and perceived depth can be modeled by the following equation:

$$D = V/(1 + \frac{I \cdot R_x}{P \cdot W}) \tag{7.4.1}$$

where D represents the perceived depth, V represents the viewing distance between observer and screen plane, I represents the interocular distance, P is the disparity in pixels, W and R_x represent, respectively, the width (in cm) and the horizontal resolution of the screen.

According to equation 7.4.1, perceived depth is not only a function of disparity but also influenced by the viewing condition, which concerns the viewing distance and the properties of display. For instance, a 8-pixel negative disparity can create a perceived depth of about 3.5 cm behind the screen when it is presented on a 24-inch full-HD stereoscopic display with 1-meter viewing distance (3 times of the screen's height). However, the same disparity corresponds to a perceived depth of infinite on a 2k cinema screen with 8-meter screen height and 8-meter viewing distance. When the viewing condition varies, the change of perceived depth from even a same disparity value might make some areas of a 3D scene impossible to fuse. Consequently, the saliency distribution can be

different. In this chapter, we adopt Equation 7.4.1 to compute the depth map for each image, the interocular distance is set to 6.3 cm, while the screen property parameters are set according to the setup of the eye-tracking experiment (introduced in Section 5).

7.4.2 A Bayesian approach of depth saliency map generation

In the area of saliency map creation, Bayes's theorem has been widely applied in various ways (e.g. [Zhang 08, Chikkerur 10, Pinneli 08]). In this chapter, we propose a new approach to apply Bayes's theorem for computing a depth saliency map based on features extracted from a depth map. The proposed approach correlates depth features with the level of depth saliency, by using the data from a psychophysical experiment.

We firstly introduce the proposed definition of depth saliency: the depth saliency (S) of each location (a pixel) equals to the probability of this point being gazed at, given depth features observed at this point and the spatial location of this point:

$$S = P(C = 1|\bar{f}_{dep}, l_z) \tag{7.4.2}$$

where C is a binary random variable denoting whether or not a point is gazed at. The random variable vector $\bar{f}_{dep}$ denotes depth features observed at this point, l_z denotes its location in depth. Note that the term about 'features', $\bar{f}_{dep}$, stands for not only the local visual features such as relative depth (i.e. disparity) and absolute depth (i.e. distance to observer), but also some higher order features considering the information from neighborhood, such as the result of applying Difference of Gaussian kernel (DoG) on feature maps.

Regarding to the right side of equation 7.4.2, $P(C = 1|\bar{f}_{dep}, l_z)$, we make assumptions that 1) the depth features of each point are independent of its distance to the viewer, and 2) $P(C = 1)$ is simply a constant. By using Bayes' rule, this probability can be thus transformed:

$$\begin{aligned} S &= \frac{P(\bar{f}_{dep}, l_z|C = 1) \cdot P(C = 1)}{P(\bar{f}_{dep}, l_z)} \\ &= \frac{P(\bar{f}_{dep}|C = 1) \cdot P(l_z|C = 1)}{P(\bar{f}_{dep}) \cdot P(l_z)} \cdot P(C = 1) \\ &= P(C = 1|\bar{f}_{dep}) \cdot P(C = 1|l_z) \cdot const. \end{aligned} \tag{7.4.3}$$

The first term in equation 7.4.3, $P(C = 1|\bar{f}_{dep})$, represents the probability of a point to be gazed at, given only the features extracted from depth information at this point. By computing this probability, the saliency map from depth channel can be obtained. The second term in this equation, $P(C = 1|l_z)$, represents the probability of a point to be gazed at given its distance to the viewer. This probability reflects observers' viewing strategy, the bias of eyes position, or the prior knowledge about at which distance potential targets are likely to appear. Compared to the well known 'center-bias' regarding to

the location prior in the viewing of 2D image [Tatler 07, Tseng 09], relatively little of this preference of observation in depth is known and studied. Recently, this preference was quantified and named as 'depth-bias' by Wang et al. in [Wang 11b]. Therefore, based on the proposed model of depth saliency, the saliency value of each point in a three dimensional scene can be considered as a combination of visual saliency from depth features and depth prior. However, studying depth-bias is not in the scope of this chapter. In the following part, we focus on the introduction of modeling $P(C = 1|\bar{f}_{dep})$, omitting the depth prior part.

By using Bayes' rule, we can get:

$$P(C = 1|\bar{f}_{dep}) = \alpha \cdot \frac{P(\bar{f}_{dep}|C = 1)}{P(\bar{f}_{dep})} \tag{7.4.4}$$

where α is a constant value representing the probability $P(C = 1)$. The function $P(C = 1|\bar{f}_{dep})$ represents how depth features observed at a point, influence the probability of the human visual system of deciding whether to fixate this point or not. This probability is proportional to the feature distribution at a gaze point, normalized by the rarity of features in the context (see equation 7.4.4). Note that the use of the likelihood, $P(\bar{f}_{dep}|C = 1)$ in the proposed approach differs from the way in which it is usually used by many models in the literature applying also Bayes's theory. We are not doing any binary classification to make a decision that a point is a fixation or not. Instead, we define the result (i.e. depth saliency map) as a distribution of probability of the points being gazed at as a function of depth features.

To achieve the computation of depth saliency map, the proposed approach consists of two stages: (1) depth feature extraction, and (2) probability distribution modeling.

7.4.2.1 Depth feature extraction

The proposed model uses depth contrast as feature for depth saliency map prediction. In most situations, depth contrast can be an efficient indicator of interesting target. For example, the HVS might consider a region protruding above a flat plane as a potential target [Potapova 11]; or might consider a hole as a place where potential target might exist.

Difference of Gaussians (DoG) filter was applied to the depth map for feature extraction (i.e. depth contrast). DoG filter has been widely used by computational models in the literature due to its resemblance to the receptive fields of neurons, and its capability of simulating the center-surround mechanism in human visual system. The DoG filters used in the proposed model were generated by:

$$f(x, y) = \frac{1}{2\pi\sigma^2} \exp(-\frac{x^2 + y^2}{2\sigma^2}) - \frac{1}{2\pi K^2\sigma^2} \exp(-\frac{x^2 + y^2}{2K^2\sigma^2}) \tag{7.4.5}$$

where (x, y) is the location in the filter. σ and K were used to control the scales of DoG and the ratio between the 'center' area and 'surround' area. In this chapter, we selected

a scale as $\sigma = 32$ pixels (approximately corresponding to 1 degree of visual angle in our experiment) and a center/surround ratio as 1/1.6 (the same value as the one used in [Zhang 08]).

7.4.2.2 Probability distribution modeling

In the proposed model, the function $P(C = 1|f_{contrast})$ models the relationship between the depth contrast of each position in a scene and the probability that this position is gazed. We propose to model this function based on a probability-learning of eye movement data collected from an eye-tracking experiment.

An important factor that can affect modeling is the stimuli used in the eye-tracking experiment. We prefer to use synthetic stimuli rather than natural content stimuli. Generally, 3D images of natural content contain not only depth but also many other features affecting eye movements. For instance, observers' attention could be affected by 2D bottom-up visual features such as color and luminance; or top-down features such as the presence of people, animals, or text; or center-bias caused by the preference of photographers to put the interesting objects close to the center of the scene. The simultaneous appearance of so many features increases the difficulty of evaluating how people's viewing behavior is actually affected by depth information. On the other hand, obtaining a precise depth map for natural content 3D image is still challenging in terms of the cost as well as the quality of the depth map.

In this study, eye movement data are obtained from an eye-tracking experiment using synthetic stimuli. These stimuli consisted of 3D scenes in which a background and some similar objects were deliberately displayed at different depth positions. The details of this experiment have been introduced in Section 6.3.

There were several advantages of using the proposed synthetic stimuli to collect data for learning the relationship between depth features and people's viewing behavior.

1. It is possible to precisely control the depth of objects and background. In other words, a precise depth map can be created for each scene. Moreover, the cost of generating synthetic images is less than the acquisition of natural images, which means that a great amount of stimuli can be taken advantage of. In this chapter, we designed a large set of 3186 synthetic images which were used for modeling the function $P(C = 1|f_{contrast})$.

2. Influence of 2D visual features on viewing behavior can be limited. In our experiment, all the objects were uniformly located, with a constant shape, size, and distance to the center of the screen. This setup enables the stimuli to get rid of as many bottom-up visual attention features as possible.

3. Influence of depth features coming from depth cues other than disparity can be limited. Disparity was the only depth cue elicited in this experiment. The reason of choosing binocular disparity is that its relationship with perceived depth can be

well modeled (as introduced in Section 7.4.1). However, for some other (monocular) depth cues, such as blur, perspective, occlusion, and so on, their influence on perceived depth is difficult to be quantitatively measured.

4. A white noise background and simple allocation of objects limit the complexity of the scenes presented to the observers at a low level, which made a shorter observation duration feasible. The viewing time of natural content images in eye-tracking experiments was generally set to 10 seconds or more. Compared to that, the viewing time in our experiment was relatively short (3 seconds for each condition). Nevertheless, it was still long enough for participants to subconsciously position their fixations on objects and explore the scene as they wanted. Hence, using these simple stimuli allowed experimenters to collect more data.

The probability distribution $P(f_{contrast})$ can be obtained based on the depth contrast maps of the synthetic stimuli. By considering the probability distribution of depth contrast at gaze points recorded during the viewing, $P(f_{contrast}|C = 1)$ can be then obtained. Therefore, the likelihood $P(C = 1|f_{contrast})$ which models the relationship between depth contrast and the probability of being fixated can be obtained by Equation 7.4.4. In Figure 7.4.1, we illustrate the resulting likelihood distribution $P(C = 1|f_{contrast})$. As seen in the figure, the saliency is not symmetrical distributed for positive and negative depth contrast values. For positive values which corresponds to regions protruding, the curve appears to be a linearly increasing line. A higher positive contrast value yields a larger chance on a fixation. In our experiment, higher positive contrast values result from larger distance between the objects and the background. For negative feature values which corresponds to the 'dents', the curve is similar to a logarithmic curve. As the absolute of the feature value increases, the chance on a fixation also increases, but at a slower rate. In our experiment, these negative values might result from the fixations that were attracted by the surface discontinuities (i.e. the edges) and thus located in a close surrounding area of the object.

For the implementation of the proposed model, the modeled $P(C = 1|f_{contrast})$ is applied on the depth feature map. By taking the depth contrast value at each pixel as input, the saliency value of each pixel in an image can be thus computed.

7.4.3 A Framework of computational model of 3D visual attention

In this section, we introduce the framework which integrates the depth saliency map with the saliency maps computed based on 2D visual features, and achieves the prediction of the final 3D saliency map. The general architecture of the proposed framework is presented in Figure 7.4.2.

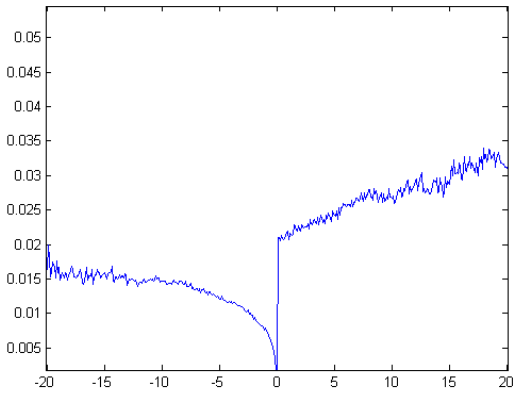


Figure 7.4.1: The distribution $P(C = 1|\bar{f}_{contrast})$ resulting from the eye-tracking experiment using synthetic stimuli.

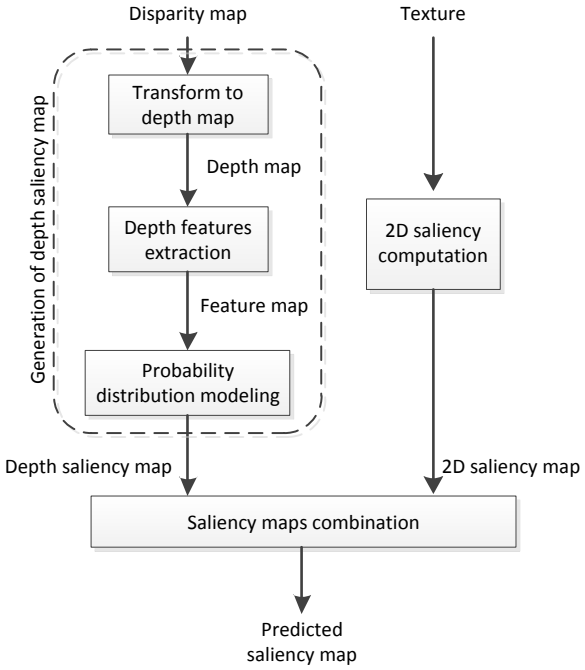


Figure 7.4.2: Overview diagram of the proposed model.

7.4.3.1 2D saliency map generation

Since developing a completely new computational model of 2D visual attention is not in the scope of the present chapter, we leave the work of 2D visual features detection and 2D saliency map creation to existing models. Three bottom-up visual attention models using quite different mechanisms were used to perform the 2D saliency prediction, and involved in the final performance evaluation:

- Itti's model [Itti 98] is one of the most widely used in the literature. The model performs a hierarchical decomposition based on three low-level visual features: luminance, color and orientation. The Matlab source code, saliencytoolbox [Walther 06], we used in this study can be downloaded on the page: `http://www.saliencytoolbox.net/`. We obtained the saliency maps by performing the 'batchSaliency' command with default parameters.

- AIM model from Bruce [Bruce 09] is based on a premise that localized saliency computation serves to maximize information sampled from one's environment. The source code used can be downloaded on the page: `http://www-sop.inria.fr/members/Neil.Bruce/`. We used the default parameters except that the rescaling factor was set to 0.25 (which means the input image was rescaled to 1/4 of its original size before the processing) to speed up the computation.

- Hou's model [Hou 07] computes the Fourier spectrum based on only luminance, and analyzes the spectral residual of an image. The source code used can be downloaded on the page: `http://www.klab.caltech.edu/~xhou/`. We used the default parameters.

In the proposed model, 2D saliency computation is only performed based on the image from the left view which is selected arbitrarily, since the images from the two views are quite similar, and the difference in 2D features between the two views' images has thus only marginal influence on visual attention deployment. Computing a 2D saliency map based on only one view instead of both views can be beneficial to decrease the computational complexity.

7.4.3.2 Saliency maps combination

The goal of this saliency maps' combination stage is to mix together the saliency maps obtained from different visual dimensions (i.e. depth information and 2D visual features). Since the 2D saliency map input is already the result of a pooling stage contained in the 2D visual attention model applied, this saliency maps combination stage focuses on merging only one 2D saliency map and the depth saliency map.

In the literature, although several approaches combining conspicuity maps of 2D visual features have been proposed, there are still not any specific and standardized approach to combine saliency maps from depth and 2D visual features. In the proposed model, we

adopt a straightforward approach which is the same as the one used in [Potapova 11] to merge the depth saliency map (SM_{dep}) and 2D saliency map (SM_{2D}): the final saliency map SM_S is equal to the sum of both maps (see Equation 7.4.6):

$$SM_S(i,j) = \omega_1 SM_{dep} + \omega_2 SM_{2D} \tag{7.4.6}$$

where $\omega_1 = \omega_2 = 0.5$.

7.5 Performance assessment

In order to assess the extent to which the depth saliency map can influence the prediction of the saliency map, and the overall performance of the proposed computational model, both qualitative and quantitative comparisons between the fixation density map and the output of the proposed model are performed in this section. All the results are obtained based on our 3D image eye-tracking database which is introduced in Chapter 5.

7.5.1 Qualitative assessment

Figure 7.5.1 gives examples of the performance of depth saliency maps (the predicted saliency maps created based on only depth map) and predicted saliency maps based on only 2D visual features (from the three 2D visual attention models introduced previously). The first three images are from the Middlebury dataset, while the others three are from IVC 3D image dataset. The fixation density map generated by eye-tracking data are also provided to be the ground-truth for the qualitative assessment. In each saliency map, brighter areas correspond to areas with higher saliency.

Qualitatively speaking, the proposed approach creates depth saliency maps that well predict salient areas in the 3D images. All the potential salient areas are depicted to be salient in the depth saliency maps. Compared to the depth saliency maps, the contribution of 2D visual features in predicting saliency of a 3D image largely depends on the model selected: Itti's model usually predicts some most salient areas, while it misses many areas of middle level saliency; Bruce's model significantly highlights the edges; Hou's model (as well as Bruce's model) is largely affected by the appearance of the large amount of texture in the background. For instance, in the image on the last row, the trees in the background are assigned high saliency by both these models.

7.5.2 Quantitative metrics of assessment

The goal of quantitative assessment is to quantify how well the proposed computational model of visual attention predicts fixation density maps coming from eye-tracking experiments. So far, there are no specific and standardized measures to compare the similarity between fixation density maps and the saliency maps created by computational models in 3D situation. Nevertheless, there exists a range of different measures

(a) (b) (c) (d) (e) (f)

Figure 7.5.1: Examples of the performance of different models including from left to right: (a) original image, (b) human saliency map, (c) the depth saliency maps created by the proposed model, and the saliency maps created by (d) Itti's model, (e) Bruce's model, (f) Hou's model.

that are widely used to perform the comparison between saliency maps for 2D content. The most common ones include: (1) correlation coefficient [Le Meur 06, Engelke 10], (2) Kullback-Leibler divergence [Le Meur 06, Bruce 09], and (3) area under the receiver operating characteristics (ROC) curve [Zhang 08, Zhao 11]. The former two are directly applicable to a comparison between a fixation density map and a predicted saliency map, whereas the area under the ROC curve (AUC) is usually deployed to compare the actual fixation (or gaze) points to a predicted saliency map. Since the disparity compensation for binocular eye-tracking data has been done during the process of fixation density map creation, the two fixation density maps from both views have been merged to one. We therefore adopt these three similarity measures to quantitatively compare a merged fixation density map and a predicted saliency map from one view.

7.5.3 Performance of depth saliency map

The creation of depth saliency map and the 2D saliency map are the two main parts of the proposed model. In order to assess the extent to which these two sources of saliency maps can predict the salient areas in a scene, the performance of the depth saliency map only is compared with the performance of the 2D saliency maps that come from three state-of-the-art 2D visual attention models, respectively.

	PLCC	KLD	AUC
DSM only	0.368	0.708	0.656
Itti's model	0.141*	2.809*	0.540*
Bruce's model	0.325	0.735	0.638
Hou's model	0.291	0.802*	0.630

Table 7.2: Performance of depth saliency map (noted as DSM in the table) only and three state-of-the-art 2D models. Note that a smaller KLD score means a better performance. * means that it is significantly different from the performance of the DSM (paired t-test, $p < 0.1$).

The results (see Table 7.2) from all the three objective metrics show that the depth saliency map has a significantly higher performance than Itti's model. Compared to Bruce's model and Hou's model, the performance of the depth saliency map is still higher, but without significant difference (except that the KL divergence value shows that the depth saliency map significantly outperforms Hou's model). These results demonstrate a great influence of the depth contrast on the distribution of visual attention in the viewing of 3D content.

7.5.4 Added value of depth saliency map

The proposed model in the present paper belongs to the category "depth-saliency model", which highlights the existence of a depth saliency map. To compare the two different ways of taking advantage of depth information, the performance of the following methods was measured and compared:

- No-depth method. This is a direct deployment of 2D computational model, no depth information is taken into account.
- Depth-weighting (DW) method. We adopt Chamaret's method [Chamaret 10], which weights each pixel in the 2D saliency map by multiplying it with the depth value of the corresponding pixel in the depth map (see Figure 7.3.3). Since we do not have the code to apply exactly the same 2D computational model used in their paper, the 2D saliency map creation part is replaced by the models of Itti, Bruce, or Hou.
- Depth-saliency (DS) method, i.e. the proposed computational model in this chapter. It creates a depth saliency map and a 2D saliency map respectively, then combines the resulting saliency map from both paths to get the final result.

The performance of these three methods is shown in Table 7.3. Large added values of the depth saliency map are demonstrated for all the three 2D visual attention models. The proposed model outperforms both the DW method and the 2D models in predicting salient areas of 3D images. As it is difficult to have an idea what is a good performance, we remind here the performance of these three state-of-the-art 2D models which has been validated on different 2D-image databases: Itti's model has a PLCC value ranging from 0.27 to 0.31 [Perreira Da Silva 10]; Bruce's model has a PLCC value ranging from 0.40 to 0.45 [Perreira Da Silva 10]; and Hou's model has an AUC value staying around 0.69 [Le Meur 10a]. Compared to the performance of these state-of-the-art 2D models on 2D content, the proposed model (DS method) is demonstrated to have a comparable level of performance on 3D content.

7.5.5 Content-based analysis

The proposed database provides 3D images of different types of natural scenes. The variation in performance of the depth saliency map and its added value to 2D models make a content-based analysis rather meaningful. For simplicity, (1) only Bruce's model is used as the reference to evaluate the performance and the added value of depth saliency map; and (2) we adopt only the PLCC scores for this content-based analysis. Bruce's model is selected since it shows a relatively good performance on various types of scenes (a PLCC value ranging from 0.40 to 0.45 on different 2D image datasets [Perreira Da Silva 10]). The results are shown in Table 7.4. In almost all cases (except image 7, "Moebius", in the Middlebury dataset), the proposed method has better results

		PLCC	KLD	AUC
Itti's model	2D model only	0.141	2.809	0.540
	DW method (Chamaret)	0.140	2.892	0.540
	DS method (Proposed)	0.356*	0.701*	0.656*
Bruce's model	2D model only	0.325	0.735	0.638
	DW method (Chamaret)	0.311	0.810	0.639
	DS method (Proposed)	0.423*	0.615	0.674
Hou's model	2D model only	0.291	0.802	0.630
	DW method (Chamaret)	0.290	0.878	0.633
	DS method (Proposed)	0.409	0.603*	0.669

Table 7.3: Contribution of the depth information on 2D models. Note that a smaller KLD score means a better performance. * means that it is significantly different from the performance of the corresponding 2D model (paired t-test, $p < 0.1$).

than the 2D saliency map. However, the depth-weighting method (a multiplication of 2D saliency and depth map), only obtains the best result for one scene. In this scene (image 12 "Hall"), all the potentially salient areas have already been detected by the 2D visual attention model.

In order to further investigate the influence of depth saliency map on the prediction of salient areas, a content-based analysis is done in terms of the performance of depth saliency map and the added value of depth saliency map. We compute the difference of PLCC value for each image by Equation 7.5.1 and Equation 7.5.2:

$$\Delta^1_{PLCC} = PLCC_{dep} - PLCC_{2D} \tag{7.5.1}$$

$$\Delta^2_{PLCC} = PLCC_{combined} - PLCC_{2D} \tag{7.5.2}$$

In Figure 7.5.2, one can observe a linear relationship between the performance of the depth saliency map and its added value. A higher performance of the depth saliency map corresponds to a higher added value. Image clustering patterns can be also clearly observed: (1) four images are located in the region of higher performance of depth saliency map and high added value, (2) two images are placed in the region of lower performance and low (or even negative) added value, and (3) the remaining twelve images are spread around a comparable performance between depth saliency map and 2D model, with a considerable added value.

When taking a closer look at the four images with high performance and high added value, one can observe that these images contain a great amount of texture or salient 2D visual features. In image 1 (Art) and image 4 (Dolls), the widespread presence of face and artificial color attracts viewer's attention to most of the areas in the scene. On the

	2D model (Bruce)	Depth saliency map	Chamaret's method	Proposed method
Image 1	0.113	**0.402**	0.042*	0.319*
Image 2	0.364	0.373	0.512*	**0.519***
Image 3	0.321	0.384*	0.231*	**0.449***
Image 4	0.240	**0.542***	0.247	0.459*
Image 5	0.252	0.114*	0.209*	**0.258**
Image 6	0.568	0.507*	0.532*	**0.595***
Image 7	**0.413**	0.198*	0.394*	0.372*
Image 8	0.447	0.390*	0.376*	**0.531***
Image 9	0.379	0.401*	0.336*	**0.454***
Image 10	0.271	0.269	0.272	**0.343***
Image 11	0.345	0.322*	0.159*	**0.413***
Image 12	0.321	0.302*	**0.439***	0.370*
Image 13	0.501	0.469*	0.272*	**0.591***
Image 14	0.344	0.462*	0.291*	**0.462***
Image 15	0.513	0.517	0.509	**0.607***
Image 16	0.232	0.225	0.232	**0.265***
Image 17	-0.134	**0.139***	-0.062*	0.013*
Image 18	0.367	0.598*	**0.603***	0.595*

Table 7.4: Performance (based on the metric PLCC) of 2D model, depth saliency map, and the added value of depth achieved by (Chamaret's) depth-weighting method and the proposed method. * means that it is significantly different from the Bruce's 2D model (paired t-test, $p < 0.1$). Note that the ID of each image is indicated in Figure 5.2.1 and Figure 5.2.2.

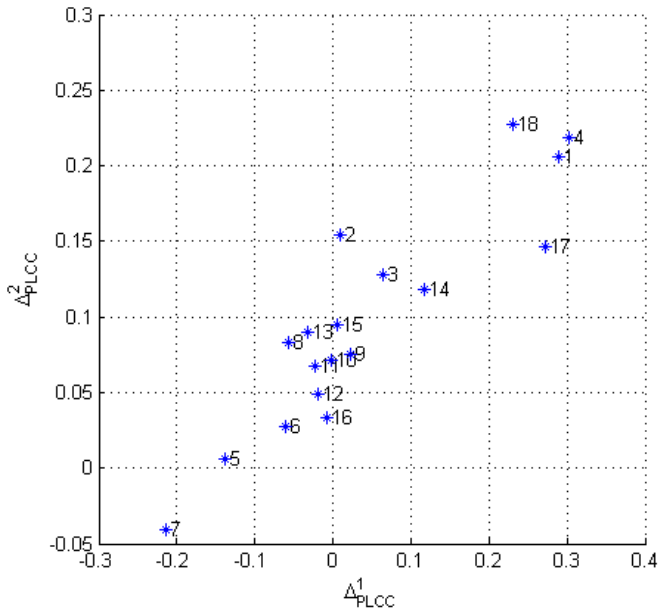


Figure 7.5.2: The scatter plot of the performance and added value of the 18 images. The ID of each image is indicated in Figure 5.2.1 and Figure 5.2.2.

other hand, in image 17 (Tree Branches), one can find a huge amount of texture with similar color and no presence of any object of interest. We can make the hypothesis that, when too much or too little area is detected as salient by 2D visual features, it would be difficult for 2D visual attention models to detect the real salient region in the scene. Depth thus become the dominant feature to direct viewer's attention. As for the image 18 "Umbrella", it represents a typical kind of 3D scene: the main actor (or object) is given a positive disparity in order to create the "popping out" effect. In this situation, observers' attention is attracted by the popping out object, and the depth feature is obviously the dominant feature directing observers' visual attention.

Image 5 (Laundry) and 7 (Moebius) are the two which yield worst performance and lowest added value of depth saliency map. In the image 5 "Laundry", the appearance of top-down features (text on the packing of the laundry detergent) attracts much attention. In the image 7 "Moebius", most of the objects of interest are with a color different from their surrounding background. These objects are located among some other objects of no-interest with the colors similar to the background. This special setup of a scene enables 2D visual attention models to detect the salient areas easily. However, in terms of depth contrast, all the objects in the scene are considered as salient. The performance of the model is thus decreased by the increasing number of "false-positive" detection. This result shows that 2D visual features should be treated as dominant features in the situations when the scene is "manually" designed to distinguish some of the objects from the others by using 2D visual features (such as color, luminance, etc).

The remaining twelve images yield a comparable performance of depth saliency maps to 2D saliency maps. However, they yield also a considerable added value. One can observe that the common characteristic of these images is that they generally have a small number of salient areas, which can be caused by either 2D salient features or depth. Therefore, the saliency map generated based on either 2D salient features or depth might predict part of the salient areas, but not all of them. This can be the reason why 2D saliency maps and depth saliency maps have comparable performance, but their combination has a much better result.

7.6 Conclusion and discussion

Studies concerning a computational model of 3D visual attention are presented in this chapter. We firstly introduce previous work on 3D visual attention, and propose a new taxonomy of computational models of 3D visual attention. After the bibliography study, we present a depth-saliency based computational model of visual attention for 3D still images.

The proposed model contains a depth saliency map creation part which is based on a probability-learning from the experimental data. By integrating the depth saliency map with the results of 2D visual feature detection, the proposed model has a good prediction of salient areas. On the other hand, it is demonstrated that the performance of depth saliency map and its added value to 2D model vary across different types of scene. When too many areas (or no area) can be indicated as salient based on 2D visual features, depth becomes the dominant cue that directs viewer's attention. But the depth saliency map does not perform well in showing the salient areas that are caused by only 2D visual features.

Two different ways of applying depth information in 3D visual attention models are compared in our study. Our results show that, creating a depth saliency map based on depth contrast achieves a higher performance than a simple depth-weighting method (a multiplication of 2D saliency map and depth map). This result indicates the importance of a depth saliency map in modeling 3D visual attention. Nevertheless, this result should not lead to a strong conclusion that a depth-saliency model is definitely better or worse than a depth-weighting model, since the depth-weighting model also has various advantages, such as the low computation complexity, or the comparable performance for some types of scenes. On the other hand, it would be reasonable to suggest that an efficient 3D visual attention model can be an integration of both types of models: depth information is treated as a additional visual dimension from which depth features are extracted to create depth saliency maps; as a possible extension, the location information (e.g. center-bias or depth-bias) is also used as weighting information to relate the distribution of attention and the distance between observer and each object in the scene.

Key points

Context

- ❏ Studies have demonstrated effects of 2D features and influences of depth information in the viewing of 3D content. Based on these conclusions, several computational models of 3D visual attention have been proposed in the literature. These models share a same step of detecting 2D saliency based on 2D visual features, and a step of integrating influences of depth information.
- ❏ The existing models can be classified into three categories: depth-weighting models, depth-saliency models, and stereo-vision models, according to the way of using depth information. In recent years, the studies of the former two types of models have received most of the attention. The key difference to identify these two categories from each other is the existence of a step of computing depth-saliency map based on depth features.
- ❏ The limitation of existing 3D visual attention models includes: (1) taking into account results of psychophysical experiments about depth's influence only in a qualitative way; (2) the lack of quantitative evaluation of performance.

Contributions

- ❏ We propose a new taxonomy of existing computational models of 3D visual attention, based on the way of using depth information.
- ❏ We propose a depth-saliency-based model of 3D visual attention. Bayes's theorem is applied on results of an eye-tracking experiment for computing the depth saliency map. Two different ways of integrating depth information in 3D visual attention model are investigated. The results demonstrates a large added value of depth saliency map and a good performance of the proposed depth-saliency model.

Chapter 8

Center-bias in stereoscopic 3D visual attention models

In the viewing of 2D images or videos, a so-called "center-bias" (or "central fixation bias") has been demonstrated: gaze fixations are biased towards the center of the scene [Tseng 09, Tatler 07]. Studies [Le Meur 06, Ma 02] have demonstrated that the performance of 2D saliency models can be largely improved by integrating the center-bias.

Therefore, one possibility to improve the performance of the 3D visual attention model is to take into account the center-bias. However any study that focuses on the approach of combining 3D visual attention models with center-bias is still lacking. Moreover, the degree to which center-bias in 3D viewing condition differs from the one in 2D viewing condition is still unknown. What is the difference of center-bias between the left and the right view also remains an open question.

In this chapter, we present a study focusing on modeling the center-bias and integrating it into the computational model. We quantitatively evaluate the center-bias for both 2D and 3D viewing, and propose an approach to combine center-bias with visual attention models for 3D viewing condition. Finally, by combining the work of this chapter and the previous chapter, we propose a hybrid 3D visual attention model which is based on 2D saliency, depth saliency and center-bias.

8.1 Introduction

It has been demonstrated that 3D visual attention, is still guided by many 2D visual features [Jansen 09]. This consistence of the influence of 2D low-level features implies the possibility of extending existing 2D models for 3D applications. This is also the reason why most of the existing computational models of 3D visual attention share a same step in which salient regions are first detected based on 2D visual features [Zhang 10, Chamaret 10, Bruce 05].

Actually, besides the salient regions resulting from 2D visual features, fixation patterns

from eye-tracking experiments have also demonstrated a bias towards screen center. This phenomenon is named as "center-bias" (or "central fixation bias"). The causes of this center-bias effect include the photographer bias, the viewing strategy, the orbital reserve, the motor-bias, and the center of screen bias [Tseng 09]. Studies have indicated that the prediction of salient region can be largely improved by integrating the center-bias effect (e.g. [Zhao 11, Ma 02, Luo 11]).

However, center-bias has not been taken into account in most of the existing 3D visual attention models. There still exist several difficulties of applying center-bias in 3D visual attention models:

1. The influence of center-bias in 3D viewing has not been confirmed. Several studies [Hakkinen 10, Ramasamy 09] draw inconsistent conclusions about how the extent of fixation distribution varies from 2D viewing to 3D viewing. The variation of extent implies the different degrees of center-bias in the two viewing conditions. Therefore, studies concerning the degree of center-bias in 3D viewing condition is indispensable before integrating the effect of center-bias into computational modeling of 3D visual attention.

2. The ways of integrating center-bias with 3D visual attention are not consistent. Not all 3D visual attention models can combine center-bias in the same way. To the models taking both views as input (e.g. [Bruce 05]), center-bias can be added on both views; to the models taking one image and a depth map as input (e.g. [Zhang 10]), center-bias has to be added as a post-processing step after the output of saliency map. This difference implies that center-bias might be integrated, at different steps, into the computational models .

3. There are still few databases providing both 2D and 3D versions of the same set of images, and the corresponding eye-tracking data. The lack of ground truth data limits the study of center-bias in 3D condition.

In next section, we propose a simple computational model of 3D visual attention which can easily take advantage of center-bias and existing 2D models. The degree of center-bias during 3D natural content images viewing has also been quantitatively evaluated. Our results indicate a clear difference between center-bias in 2D and 3D. By using proper center-bias in the proposed model, a significant added value of center-bias has been demonstrated in the prediction of saliency maps for 3D images.

8.2 A simple 3D visual attention model

The proposed model is inspired by an attentional framework for stereo vision proposed by Bruce and Tsotsos [Bruce 05]. This attentional framework was selected on the basis of its biological plausibility. We simplify this framework due to its high level of complexity. The simplification was achieved by keeping only layer 1 which corresponds to

the detection of salient areas based on 2D visual features, and layer 2 which corresponds to a shift in attention according to various binocular disparities.

In the proposed model, the left-view image and the right-view image are taken individually as the inputs. Firstly, a 2D visual attention model is applied independently on the two images, and creates a corresponding 2D saliency map for each view. Secondly, the left and the right 2D saliency maps then go through an attention shifting step in which two saliency maps are merged. In this model, center-bias can be either added in both paths to weight the two 2D saliency maps before the attention shifting step (Figure 8.2.1 right); or be added after this step to weight the fused saliency map (Figure 8.2.1 left). The details of these various steps are introduced in this section.

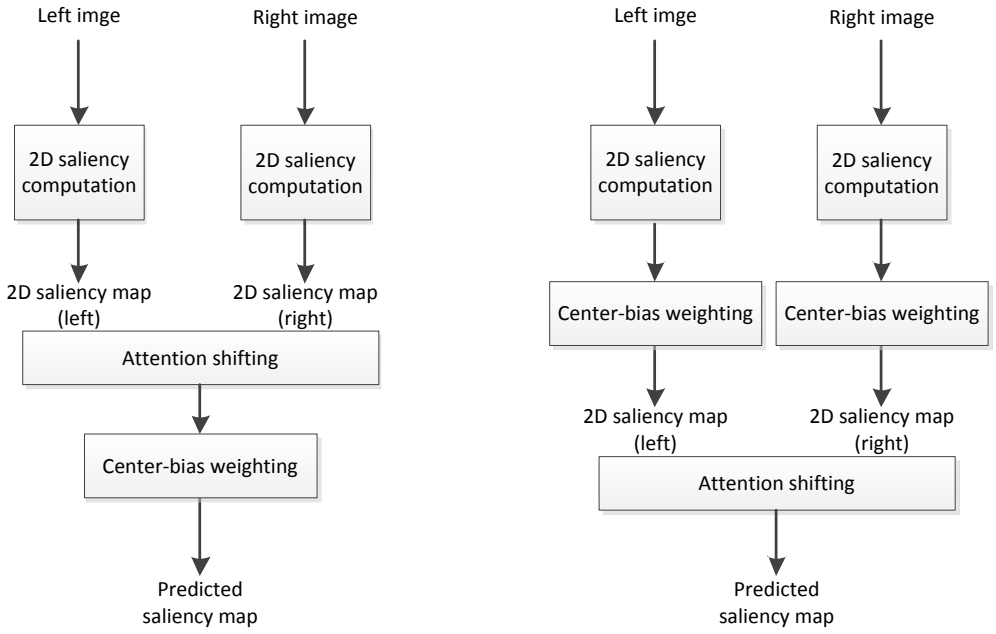


Figure 8.2.1: Overview diagrams of the proposed model with two ways of integrating center-bias.

8.2.1 2D saliency computation

Since developing a completely new computational model of 2D visual attention is not in the scope of this study, we leave the work of 2D visual features detection and 2D saliency map creation to existing models. In this study, three widely used models using different mechanisms have been evaluated:

1. Bruce et al.'s AIM model [Bruce 09] which is based on information maximization.
2. Itti's model [Itti 98] which is the most widely used one in the literature. This model is based on three low-level features, including intensity, color and orientation.

3. Hou's model [Hou 07] based on the computation of spectral residual.

Each of these three models is used to perform 2D saliency prediction in the proposed 3D model.

8.2.2 Center-bias modeling

There is still not a strong agreement on the ways of modeling the center-bias. In the literature, the center-bias was usually modeled by using either an isotropic Gaussian kernel ([Zhao 11]) or an anisotropic Gaussian kernel ([Le Meur 06]). Tseng et al. [Tseng 09] have demonstrated that image's borders have a large impact on the center-bias. It implies that the shape (i.e. the length to width ratio) of the image should be also taken into account when designing the Gaussian kernel.

In our study, the center-bias used for weighting a saliency map S is modeled by an anisotropic Gaussian kernel located at the center of image [Le Meur 06]. The weighted saliency map, named S', is then given by:

$$S'(x,y) = S(x,y)exp\left(-\frac{(x-x_0)^2}{2\sigma_x^2} - \frac{(y-y_0)^2}{2\sigma_y^2}\right)$$

where (x_0,y_0) represent the image's center coordinates. σ_x and σ_y denote the standard deviation related to the x-axis and the y-axis, respectively.

The relationship between σ_x and σ_y is quantified according to the size of image viewed:

$$\sigma_y = \sigma_x \times \left(\frac{R_x}{R_y} Ind(R_x < R_y) + \frac{R_y}{R_x} Ind(R_x > R_y)\right)$$

where R_x and R_y are the image's width and height, and $Ind()$ is the indicatric function. Note that the standard deviations σ_x and σ_y, representing the degree of center-bias are measured in visual degree, since the measurement of visual degree takes into account the viewing distance.

8.2.3 Attention shifting

Due to the disparity between the left view and the right view, an area in the scene can thus correspond to slightly different locations in retinal images of the two eyes. Moreover, since conflicts may exist between the two eyes due to occlusions in binocular viewing, the saliency maps of the left view and the right view may not be necessarily the same for all the locations in the scene. Consequently, the two saliency maps that come from the two eyes need to be merged by shifting each pixel's saliency value from one view to the other.

The distance of shifting is processed according to the local disparity between the two views. Due to the symmetry of binocular disparity, saliency map from either of the two views can be shifted to fit the other one. We thus arbitrarily shift the saliency map

of the right view, and then combine it with the saliency map of left view. The result saliency map S" is obtained by Equation 8.2.1:

$$S"(i,j) = S_L(i,j) + S_R(i + D_x(i,j), j + D_y(i,j)) \tag{8.2.1}$$

where (i,j) represents the coordinate of each pixel in the image; S_L denotes the left-view saliency map; S_R denotes the right-view saliency map; D_x and D_y denote the horizontal and vertical disparity at each pixel.

8.2.4 Result and analyses

In this section, the analyses are based on our eye-tracking database which is introduced in Chapter 5.

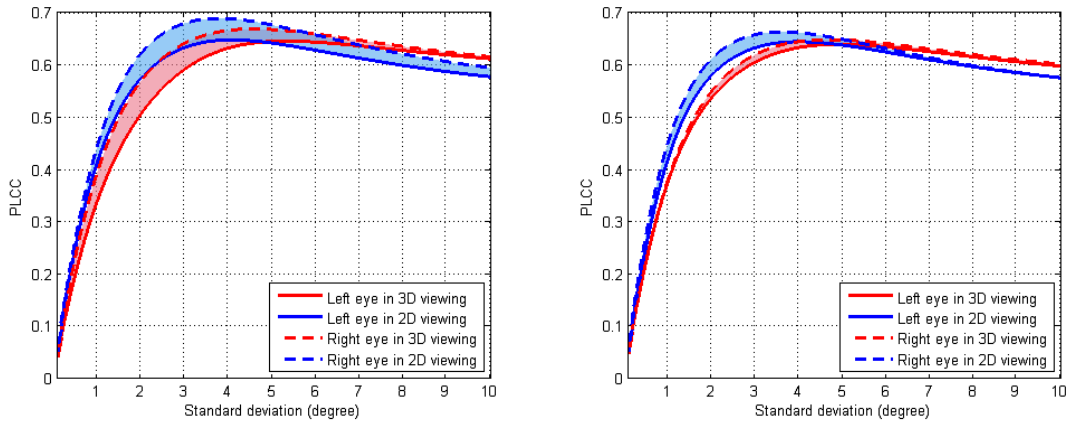
8.2.4.1 Qualitative analysis of the center-bias in 2D viewing and 3D viewing

Figure 8.2.2 shows some examples of the fixation density maps. They are obtained during the viewing of the 2D version and the 3D version of the same set of images. From the fixation density maps, clear difference of fixation distribution can be observed. The fixations are more widely distributed in the 3D images than in the 2D images.

Figure 8.2.2: Examples of fixation distribution: (a) Original image; and fixation density maps from the viewing in (b) 2D condition, (c) 3D condition.

8.2.4.2 Quantitative analysis of the center-bias in 2D viewing and 3D viewing

To quantitatively examine the degree of center-bias in 2D viewing and 3D viewing, we apply a similar method as the one used in [Le Meur 06]. A set of center-bias maps are firstly created by using only the center-bias model introduced in section 8.2.2 with a standard deviation σ_x ranging from 0 degree to 10 degree. Average Pearson linear correlation coefficients (PLCC) are then computed between each fixation density map and a set of center-bias maps. Each of these center-bias maps is compared with left and right fixation density maps of each image (both 2D and 3D version). The average PLCC evolution is plotted in Figure 8.2.3.



(a) (b)

Figure 8.2.3: The left (resp. right) figure are obtained from subjects who are left-eye (resp. right-eye) dominant. Solid lines represent the degree of center-bias of right eye's data. Dash lines represent the degree of center-bias of left eye's data. Blue color and red color are, respectively, assigned to the lines and the areas between them in order to distinguish the 2D and 3D viewing conditions.

In Figure 8.2.3, the value of PLCC represents the similarity between the real fixation distribution and a 2D Gaussian distribution; the value of standard deviation represents how much the Gaussian distribution is concentrated to the center. Therefore, a smaller standard deviation and a higher PLCC value (1) mean that the distribution of fixation is more concentrated at the center, and (2) thus correspond to a higher degree of center-bias.

From the two colors in Figures 8.2.3, it can be clearly observed that there exists a higher degree of center-bias (of both eyes) in 2D viewing than in 3D viewing. Fixations are with a generally wider distribution in the viewing of 3D content. This conclusion holds for (1) both left eye and right eye, and (2) for both the left-eye dominant viewers

and the right-eye dominant viewers. Our finding is consistent with the results from Hakkinen et al. [Hakkinen 10] and Jansen et al. [Jansen 09]. Regarding the development of computational models, a lower degree of center-bias in 3D viewing means that the Gaussian kernel applied in the 3D visual attention model should have a larger standard deviation.

Although the observers involved in our experiment have different dominant eyes, the relative difference between 2D viewing and 3D viewing existing in these two groups of observers are similar. For simplicity, we merge the data from these two groups of observers in the following analysis.

8.2.4.3 Performance of the proposed model and added value of center-bias in 3D visual attention models

Curves in Figure 8.2.3 also show differences of center-bias between left eye and right eye. These differences indicate the plausibility of adding center-bias weighting on both views' saliency maps by two different Gaussian kernels. Several objective metrics are thus applied to assess the performance of such a strategy.

However, the accuracy of the final maps shows only marginal difference among three ways of applying Gaussian kernels:

- We use the model showed in Figure 8.2.1 right. Two different Gaussian kernels are applied on the saliency maps from the two views.
- We use the model showed in Figure 8.2.1 right. Two identical Gaussian kernels are applied on the left and right saliency maps.
- We use the model showed in Figure 8.2.1 left. A Gaussian kernel is applied after the fusion of the two views.

The reason of this similarity among different ways of applying center-bias could be due to the rarity of occlusion and areas with extreme disparities in the images of our database. In the following quantitative analysis, only the way of adding center-bias after the fusion of two views is used for performance assessment.

Table 8.1 gives the performance of the proposed 3D model. Each of the three 2D models is combined with various levels of center-bias to predict the saliency maps of 3D images. Three degrees of center-bias are tested:

1. Zero center-bias. In this case, no center-bias is considered. The saliency map is uniformly weighted.
2. The 2D optimal value $\sigma_1 = 4.1$ degrees. This value of standard deviation results from a training based on the fixation patterns obtained in 2D viewing condition. This smaller value corresponds to a more concentrated distribution of fixations. Equally, it means a higher degree of center-bias.

3. The 3D optimal value $\sigma_2 = 4.9$ degrees. This value of standard deviation results from a training based on the fixation patterns obtained in 3D viewing condition. This larger value indicates a wider spread distribution of fixations and a lower degree of center-bias.

The performance of each 2D model in predicting saliency maps of 2D version images is also presented in Table 8.1 as a reference. Three widely used objective metrics are applied in the performance assessment:

- Pearson linear correlation coefficients (PLCC);
- Kullback-Leibler divergence (KLD);
- Area under the ROC curves (AUC).

2D saliency model	Image	Degree of CB	PLCC	KLD	AUC
Bruce's model	2D	No center-bias	0.2853	0.8135	0.6423
	3D	No center-bias	0.3423	0.5159	0.6397
		σ_1	0.6717	0.4675	0.7358
		σ_2	0.6913	0.3532	0.7377
Itti's model	2D	No center-bias	0.1370	2.8072	0.5480
	3D	No center-bias	0.1568	2.3740	0.5483
		σ_1	0.2147	2.4450	0.5514
		σ_2	0.2165	2.3438	0.5516
Hou's model	2D	No center-bias	0.2628	0.8576	0.6386
	3D	No center-bias	0.3003	0.5805	0.6273
		σ_1	0.6120	0.5738	0.7326
		σ_2	0.6232	0.4543	0.7323

Table 8.1: Performance of the proposed model on 3D images with different 2D saliency detection algorithms and different degrees of center-bias (noted as CB in the table). The performance of these 2D attention models on 2D images is also presented.

Table 8.1 demonstrates that the performance of the proposed 3D visual attention model largely depends on the 2D saliency model adopted no matter whether center-bias is added. When no center-bias is taken into account, the proposed model generally has a comparable performance as the performance of the corresponding 2D model on 2D images. When the center-bias is considered, all these three metrics indicate great improvements of the performance of all the three 2D saliency models. Table 8.1 demonstrates also the impact of the parameter σ on the contribution of the center-bias. The proposed model has a better performance when using the parameter σ particularly tuned for 3D condition.

8.3 A hybrid model of 3D visual attention

Both depth saliency maps and center-bias have been demonstrated to have large added value in predicting the saliency maps of 3D images. We have already introduced the methods of individually integrating either depth saliency map or center-bias with 2D saliency map. In this section, we propose a new framework of 3D visual attention model, which integrates both depth saliency map and center-bias with 2D saliency maps to achieve the prediction of the final 3D saliency map.

8.3.1 A framework of integrating 2D saliency map, depth saliency map and center-bias

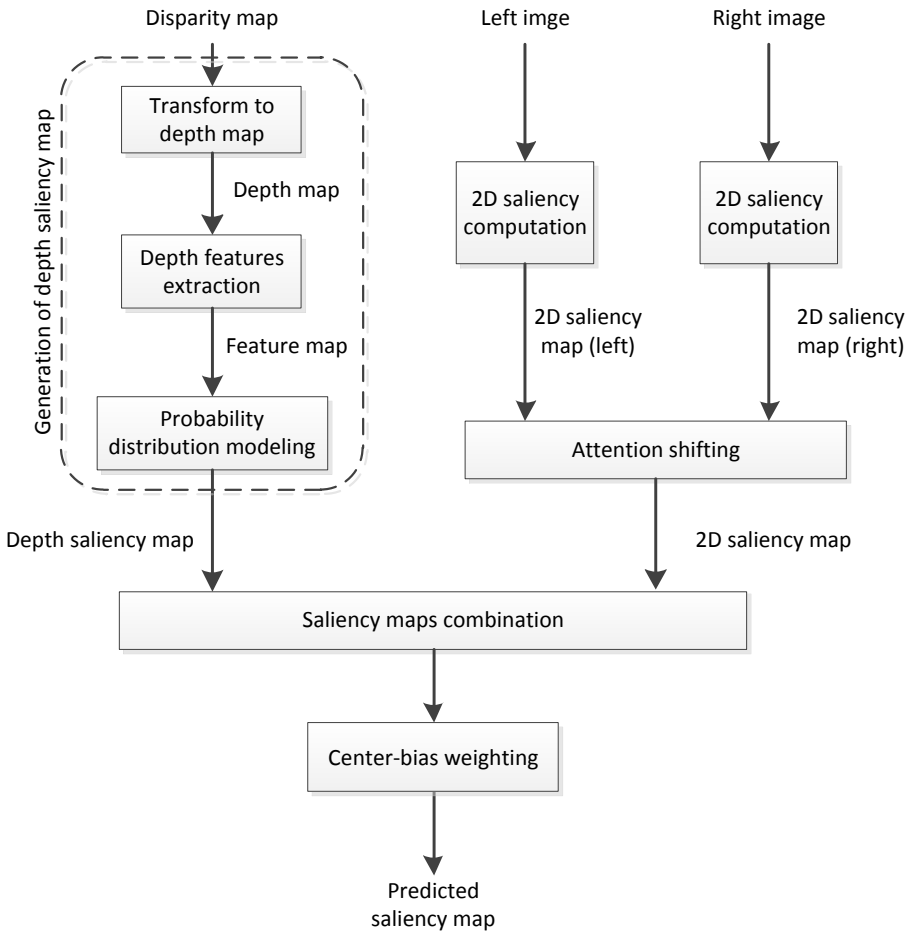


Figure 8.3.1: Overview diagram of the proposed 3D visual attention model considering both depth saliency map and center-bias.

The architecture of the proposed hybrid framework is presented in Figure 8.3.1. In

the proposed hybrid model, the left-view image, the right-view image and the disparity map are taken as input. The disparity map is used to compute the depth saliency map by following the way which is introduced in Section 7.4.2. A 2D visual attention model is used to compute the left and the right 2D saliency maps, which are then combined by an attention-shifting step which is introduced in 8.2.3. Note that only Bruce's model is tried for the purpose of computing 2D saliency maps. Bruce's model is selected here since it shows a relative good performance on various types of scenes. Next, the resulting 2D saliency map and depth saliency map are combined in a linear way as introduced in Section 7.4.3.2, and the result is then weighted by a two-dimension Gaussian kernel ($\sigma = 4.9$ degrees) representing the center-bias to get the final 3D saliency map (see Figure 8.3.2 and Figure 8.3.3).

8.3.2 Results and analyses

	PLCC	KLD	AUC
Bruce's (2D) model	0.339*	0.552*	0.640*
Bruce's (2D) model after attention shifting	0.342*	0.516*	0.640*
DSM	0.371*	0.560*	0.648*
Center-bias	0.690	0.320	0.727
Proposed hybrid model	0.726	0.322	0.744
Random	0.001*	1.153*	0.500*

Table 8.2: Performance of the proposed visual attention model. * means that it is statistically different from the performance of the proposed model (paired t-test, $p < 0.05$).

The FDM of 3D images (see Chapter 5) are used for performance evaluation. The accuracy of the proposed hybrid model's output is compared with the accuracy of another five types of saliency maps, including (1) the 2D saliency maps (obtained by Bruce's 2D model), (2) the 2D saliency maps after the "attention shifting" step, (3) the depth saliency maps, (4) the center-bias maps, and (5) the random saliency maps. Note that we create random saliency map following the same procedure as introduced in [Le Meur 10a].

Table 8.2 illustrates the results. From a quantitative point of view, the proposed hybrid model which is based on the computation of 2D saliency, depth-saliency and center-bias has a good performance. In terms of individual components, it is interesting to note that the center-bias is a better predictor than other 2D visual features and depth features. Center-bias maps outperform the 2D saliency maps and depth saliency maps whatever the metric. The overall performance of the proposed hybrid model on 3D image is already comparable (or even higher) when comparing with several state-of-the-art 2D models' performance on 2D still images. As a reference, we remind here

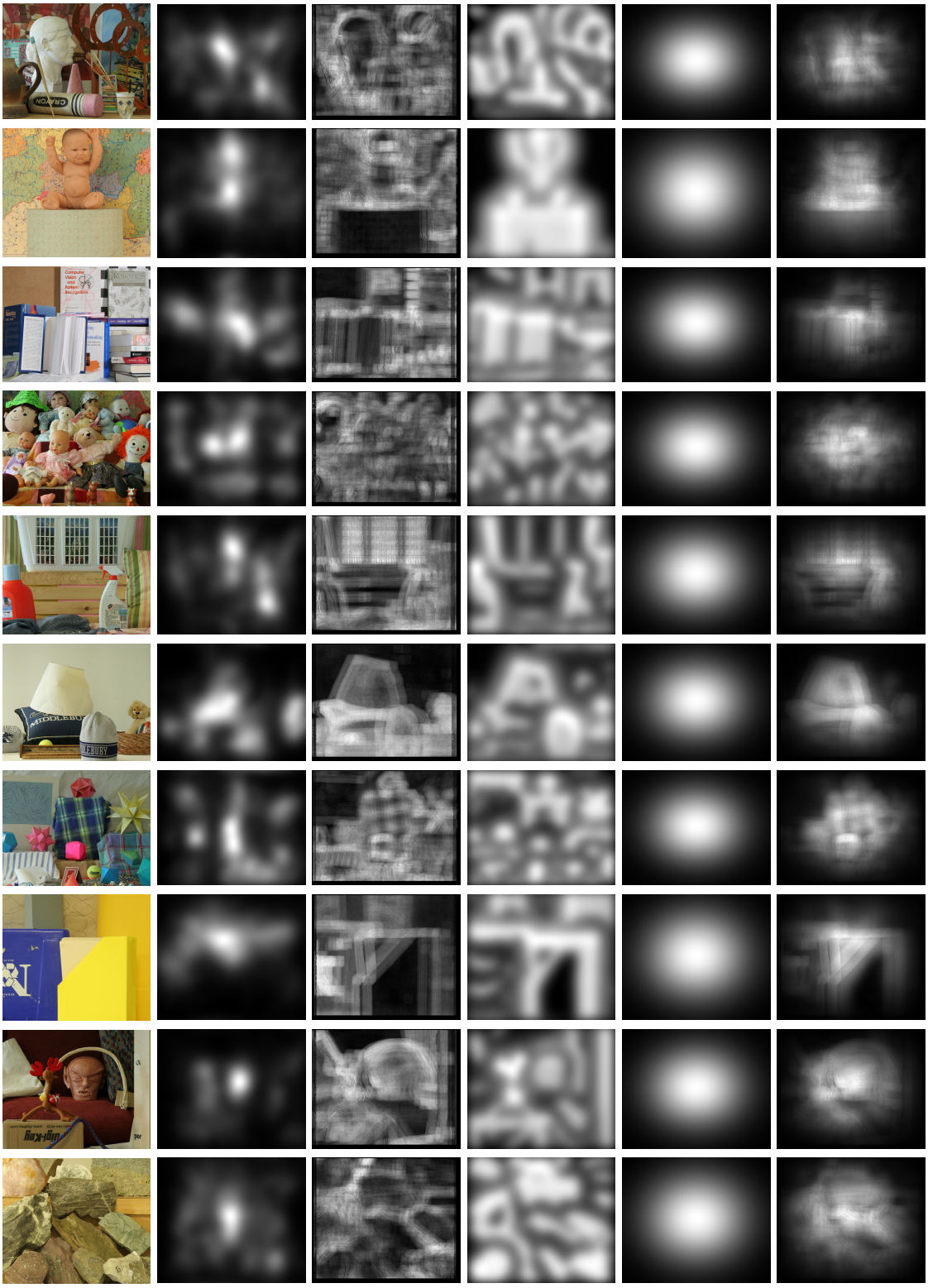
the 2D models' performance which has been validated on different 2D-image databases: Itti's model has a PLCC value ranging from 0.27 to 0.31 [Perreira Da Silva 10]; Bruce's model has a PLCC value ranging from 0.40 to 0.45 and a KLD value ranging from 1.08 to 1.59 [Perreira Da Silva 10]; and Hou's model has an AUC value staying around 0.69 [Le Meur 10a].

8.4 Conclusion

Studies concerning the integration of center-bias with 3D visual attention are presented in this chapter. A psychophysical study focusing on modeling center-bias for 3D viewing conditions is firstly presented. A 3D visual attention model to integrate the center-bias into 3D attention model is also proposed.

The proposed center-bias-based model can (1) exploit existing 2D saliency models by an attention-shifting step to combine the 2D saliency maps from both views based on the disparity maps, and (2) integrate the center-bias which has a large added value in predicting the saliency map. A good performance of this center-bias-based model has been demonstrated. On the other hand, we evaluate the difference of center-bias in 2D viewing condition and 3D viewing condition. We find out that, during watching 3D images, fixations are more widely spread in the scene. This result implies a lower degree of center-bias. As a consequence, a higher value of standard derivation should be chosen in modeling center-bias for 3D viewing condition. Moreover, our results demonstrate that the difference of center-bias between the two views has only marginal influence on the prediction of the saliency map.

Based on the center-bias-based model and the depth-saliency model proposed in the previous chapter, we finally proposed a hybrid 3D visual attention model integrating 2D saliency, depth saliency and center-bias. Both qualitative and quantitative assessments demonstrate a good performance of the proposed hybrid model on 3D images.



(a) (b) (c) (d) (e) (f)

Figure 8.3.2: Examples of the performance of different models for image 1 to 10. From left to right: (a) original image; (b) human saliency map; (c) 2D saliency maps created by Bruce's model and the "attention shifting" step; (d)depth saliency maps; (e) center-bias maps; (f) the proposed hybrid model of this section.

(a) (b) (c) (d) (e) (f)

Figure 8.3.3: Examples of the performance of different models for image 11 to 18. From left to right: (a) original image; (b) human saliency map; (c) 2D saliency maps created by Bruce's model and the "attention shifting" step; (d)depth saliency maps; (e) center-bias maps; (f) the proposed hybrid model of this section.

Key points

Context

- ❏ Center-bias has been investigated for 2D viewing condition, it has also been demonstrated to have a large added value in predicting 2D saliency map.
- ❏ It remains an open question that how center-bias varies and how it should be integrated in modeling 3D visual attention.

Contributions

- ❏ We study the variation of center-bias from 2D viewing to 3D viewing, and the way of integrating the center-bias into a 3D visual attention model. We propose a 3D model which (1) exploits existing 2D visual attention model by an "attention-shifting" process and (2) integrates center-bias which is tuned for 3D viewing condition and thus has a large added value in predicting saliency map.
- ❏ We also propose a hybrid model integrating 2D saliency, depth saliency and center-bias for predicting the saliency map of 3D image. This hybrid model has a good performance, which is comparable or even higher compared to state-of-the-art 2D models' performance on 2D images.

Chapter 9

Stereoscopic 3D visual attention and visual comfort: quantifying how the combination of blur and disparity affects the perceived depth

Quality of Experience (QoE), which is closely related to visual discomfort and visual fatigue, is an important issue that 3D-TV nowadays needs to deal with. Studies [Huynh-Thu 11a, Semmlow 79, Talmi 99] have shown that 3D visual attention information might be applied to improve QoE by collaborating with blur. In this application, 3D visual attention is used to determine the fixation point and make sure that blur is added to non-fixated areas. However, the impact of depth cues integration between disparity and blur still remains an open question.

In this chapter, we present a study focusing on quantifying how the combination of two depth cues, disparity and blur, affects depth perception in a 3D scene. Moreover, based on the assumption that fixated area is known (or it can be well predicted by visual attention models), we propose a new way of improving QoE of 3D-TV by using blur.

9.1 Introduction

Displays nowadays used to show the 3D productions are usually planar. Consequently, binocular disparity on the display plane becomes a pre-eminent depth cue which enables viewers to perceive depth.

As introduced in the previous section, binocular disparity induces a conflict between accommodation and vergence of the eyes. This conflict is usually considered as a main reason for visual discomfort, especially when the disparity is large. Therefore, the disparities of all the locations in a scene need to be limited to a certain range in order to

avoid visual discomfort. This restriction of disparity largely limits the presentation of depth information of a scene.

To solve the problem caused by the conflict between gaining apparent depth and avoiding visual discomfort (caused by disparity), a possible way is to decrease the binocular disparity of its 3D presentation, and to reinforce monocular cues to compensate the loss of perceived depth, so an unaltered apparent depth is kept.

As introduced previously, several monocular depth cues affect the depth perception. Among these monocular depth cues, the blur in the retinal image, which is induced by the limitation of depth-of-field of human eyes, is known as an important monocular depth cue. However, there is still not a strong agreement on how defocus blur influences perceived depth when it interacts with binocular disparity. Some previous investigations have shown clear contributions of blur to depth perception [Pentland 87, Watt 05, Hoffman 10, Held 10], while others showed that blur has either no effect or only some qualitative effects on perceived depth ordering [Marshall 96, Mather 96, Palmer 08].

Most of these previous studies measure the influence of the blur cue by adding blurriness only at the screen plane. However, the magnitude of each depth cue varies as a function of distance from the objects to the observer. Another limitation in most of the previous studies was the experiment apparatus. CRT monitors were used and this type of display has some disadvantages: (1) the surface displaying stimuli was slightly curved, (2) the stimuli's virtual distance was affected by refraction due to the front glass plate, and (3) the screen was usually not large enough to cover a favorable field-of-view.

In our study of quantifying how the combination of blur and disparity affects the perceived depth, we conducted a subjective experiment using a state-of-art stereoscopic display system, the Samsung SyncMaster 2233RZ which is a 22.5-inch 1680*1050@120Hz wide-screen LCD monitor working with active shutter glasses from NVIDIA. The stimuli used in the experiment contained a background plane and a single object in the foreground, both of which were chosen closer to natural content compared to the stimuli used in previous publications. The image of a butterfly was used as the foreground object, since it is spatially complex enough, containing regions with both low and high frequencies.

In our experimental observations, observers viewed stimuli in a two alternative forced choice (2AFC) task, being required to select the stimulus with largest depth interval between foreground and background. Two sources of perceived depth are used: disparity and blur. The perceived depth from disparity stems from the difference of disparity between the foreground object and the background. The perceived depth from blur stems from the amount of blur introduced to the background by convolution with a Gaussian kernel. Both the absolute position and relative distance between the foreground and background remain free parameters. This setup is able to evaluate how the combination of disparity and blur affects the perceived depth of objects located at different distances.

In Section 9.2, we introduce first the geometry about how binocular disparity and the defocus blur are created. Section 9.3 describes the experimental methods. Section 9.4 presents the results and the analysis of the experiments. Some general conclusions are

presented in Section 9.5 and Section 9.6.

9.2 Disparity and defocus blur

When people fixate an object in a three-dimensional scene, they can simultaneously perceive defocus blur and binocular disparity. Actually, the creations of defocus blur and binocular disparity in the retinal image have the same fundamental geometry.

Geometry of defocus blur

An ideal thin lens focuses parallel rays to a point on the opposite side of the lens. The distance between this point and the lens is the focal length, f. Light rays emanating from a distance d_1 in front of the lens will be focused to another point at distance s_1 at the opposite side of the lens. The relationship between these distances is derived from the thin-lens equation (see Equation 9.2.1).

$$\frac{1}{s_1} + \frac{1}{d_1} = \frac{1}{f} \tag{9.2.1}$$

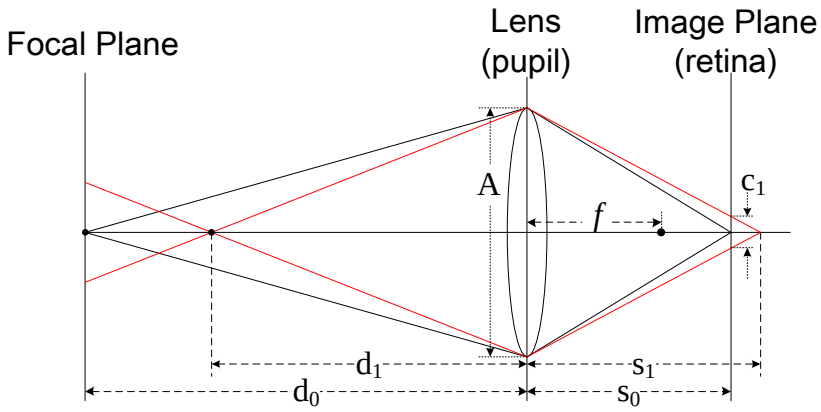


Figure 9.2.1: Schematic diagram of the generation of defocus blur.

When the lens focuses at an object at distance d_0, the image of this object will be formed at the image plane at distance s_0. This object is now in focus, while objects at other distances become out of focus and hence generate blurred images on the image plane (see Figure 9.2.1).

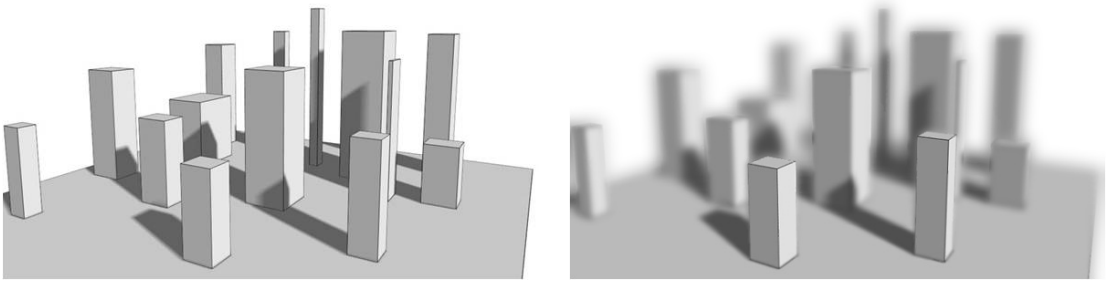
We express the amount of blur in the image plane by the blur circle diameter c_1, which can be computed by the equation 9.2.2.

$$c_1 = |A \cdot \frac{s_0}{d_0} \cdot (1 - \frac{d_0}{d_1})| \tag{9.2.2}$$

where A is the diameter of the lens (pupil) and s_0 is the posterior nodal distance (approximately the distance from the pupil to the retina). Note that the human eyes have imperfect optics, and Equation 9.2.2 does not incorporate any of the eye's aberrations. Nevertheless, previous researches showed that Equation 9.2.2 can provide an accurate enough approximation of blur when the eye is defocused [Cheng 04].

Based on this calculation, blur of the retinal image can be thus modeled as a two-dimensional Gaussian function. The blur circle radius (c_1) equals the radius of decay to $\exp(-0.5)$, i.e. the standard deviation sigma of the Gaussian function. It is thus possible to simulate the defocus blur on an image by convolving the image with a 2D Gaussian kernel.

Therefore, given A and S_0, the magnitude of blur varies according to the focus distance and the distance between the object focused and the object defocused. The magnitude of blur increases as the distance from the fixated object increases. When an object is located outside of the depth of field, this object will be projected in the eye with a perceivable magnitude of blur. This relation enables blur as a cue of depth. An example is shown in Figure 9.2.2.



(a) (b)

Figure 9.2.2: Effect of blur on perceived depth. No blur is introduced in the image (a), while blur is added to the image (b) supposing that the top of the nearest box is fixated.

Geometry of binocular disparity

Based on the same geometry (Figure 9.2.3), disparity provides also depth information to the human brain. When the two eyes separated by a distance p converge on an object at distance d_0, another object at distance d_1 creates images with an angular disparity δ which can be obtained by the following equation using small-angle approximation (see Equation 9.2.3).

$$\delta = \phi_L - \phi_R = 2\left[tan^{-1}\left(\frac{p}{2d_1}\right) - tan^{-1}\left(\frac{p}{2d_0}\right)\right] \approx p \cdot \left(\frac{1}{d_0} - \frac{1}{d_1}\right) \tag{9.2.3}$$

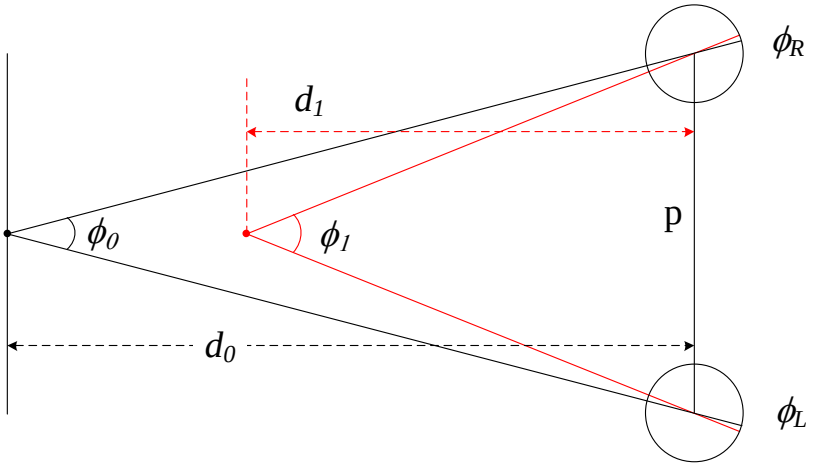


Figure 9.2.3: Schematic diagram of the generation of binocular disparity.

We can consider that disparity is caused by the difference in vantage point of two cameras (eyes), while blur is caused by the difference in vantage point at two positions of one camera (eye). By converting the equation of blur circle radius into angular units, and rearranging both equations, we can find the relationship between the magnitudes of disparity and blur (see Equation 9.2.4).

$$c_1 = \frac{A}{p}|\delta| \tag{9.2.4}$$

Therefore, the magnitudes of blur and disparity created by an object in a three-dimensional scene are shown to be proportional to each other. This relationship generally holds because accommodation and vergence are coupled in the real world, namely the eyes accommodate and converge to the same distance. It suggests that the human visual system might use information from both factors in correlating the two eyes' images. However, when people watch 3D images or videos displayed on a planar stereoscopic display, there exists a conflict between accommodation and vergence. No matter how the vergence changes, the image is always perceived sharpest when the eyes accommodate on the screen plane. This conflict suggests that a manual creation of blur on the focal plane (the screen) might inhibit the conflict and affect the perceived depth.

9.3 Experiment

In our study, a psychophysical experiment was conducted to quantify how the combination of blur and disparity affects the perceived depth.

Subjects

Thirty-five subjects participated in the experiment. Twelve subjects are male, twenty-three are female. The subjects did not know about the purpose of the experiment. Subjects ranged in age from 17 to 40 years. All the subjects had either normal or corrected-to-normal visual acuity, which was verified by three pretests before the start of eye-tracking experiment:

1. Monoyer chart test was performed to check the acuity (subject must get results higher than 9/10);
2. Ishihara test was performed to check color vision (subject should be without any color trouble);
3. Randot stereo test was performed to check the 3D acuity (subject should get results higher than 7/10).

Apparatus and stimuli

Stimuli were displayed on a Samsung 22.5-inch LCD screen (1680 × 1050 at 120 Hz). Each screen pixel subtended 65.32 arcsec. The display yielded a maximum luminance of about 50 cd/m^2 when watched through the activated shutter glasses. Stimuli were viewed binocularly through the active shutter glasses (NVIDIA 3D Vision kit) at a distance of approximately 90 cm. The peripheral environment luminance was adjusted to about 44 cd/m^2. When seen through the eye-glasses, this value corresponded to about 7.5 cd/m^2 and thus to 15% of the screen's maximum brightness as specified by ITU-R BT.500.

Each stimulus consisted of one single object in the foreground and a background. A 400 × 400 pixels butterfly image was used as the foreground object. This stimulus was easy to accommodate because it was spatially complex and therefore contained a wide range of spatial frequencies from low to high. The butterfly was shown with a horizontal offset between the left and the right view in order to create a disparity cue. The magnitude and direction of this offset varied to supply a variety of near or far disparity. At the 90 cm viewing distance, the background plane subtended 29.8 × 18.9 arcdeg, and the foreground object subtended 7.2 arcdeg.

The background plane was a photo of a flowerbed. The background looked natural. It contained a great amount of textures, contained no distinct region of interest distracting the observers. When required, the background was spatially blurred by applying a Gaussian blur kernel with a 5-pixel radius. This amount of blur equals the blur created by supposing that the focused object (foreground) and the defocused object (background) are at a distance of 13.7 cm in front of the screen and 19.8 cm behind the screen respectively, both of which are the limitations of the comfortable viewing zone derived from the 90 cm viewing distance. Note that all the blurred backgrounds were with the same amount of blur.

Design and procedure

In each trial, a pair of stimuli were shown to the subjects. One stimulus contained a blurred background (denoted as BB-stimulus) and a sharp foreground object, while the other stimulus contained a sharp background (denoted as SB-stimulus) and also a sharp foreground object (see Figure 9.3.1).

The backgrounds of both stimuli were positioned at the same depth. Two parameters of the stimuli were varied: the absolute position (D_a) of the background plane and the relative distance (D_r) of the foreground object to the background. The selection of absolute distance ranged from -19.7 cm to 6.6 cm in steps of 6.6 cm (negative values denote the positions behind the screen plane, and positive values denote the positions in front of the screen plane). The relative distance ranged from 0 cm to approximately 33 cm. All the positions of both background and foreground objects are selected considering the limitations of the comfortable viewing zone [Chen 10]. There are thus twenty combinations of absolute position and relative distance in total for the BB-stimuli.

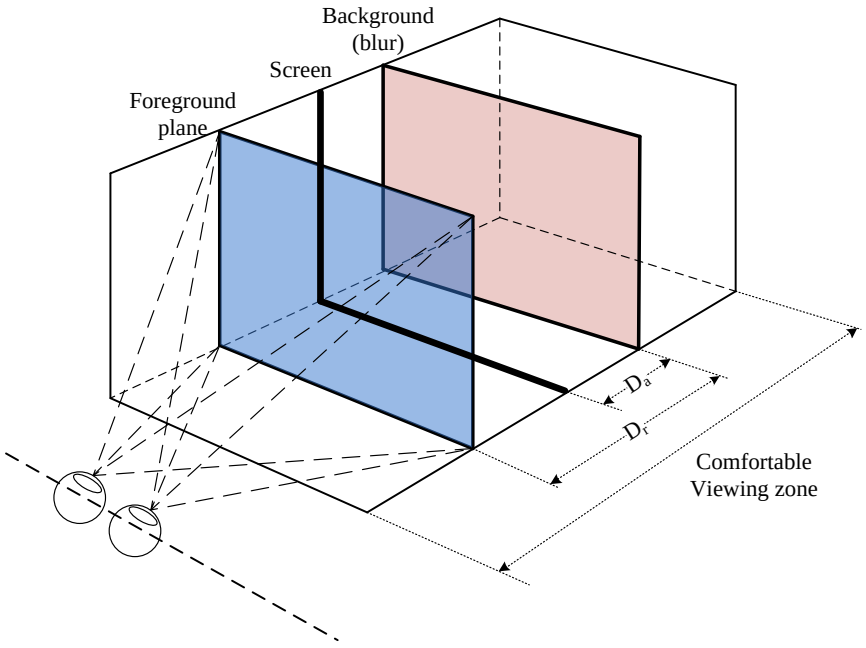
(a) (b)

Figure 9.3.1: Example of stimuli: the left views of an SB-stimulus and a BB-stimulus.

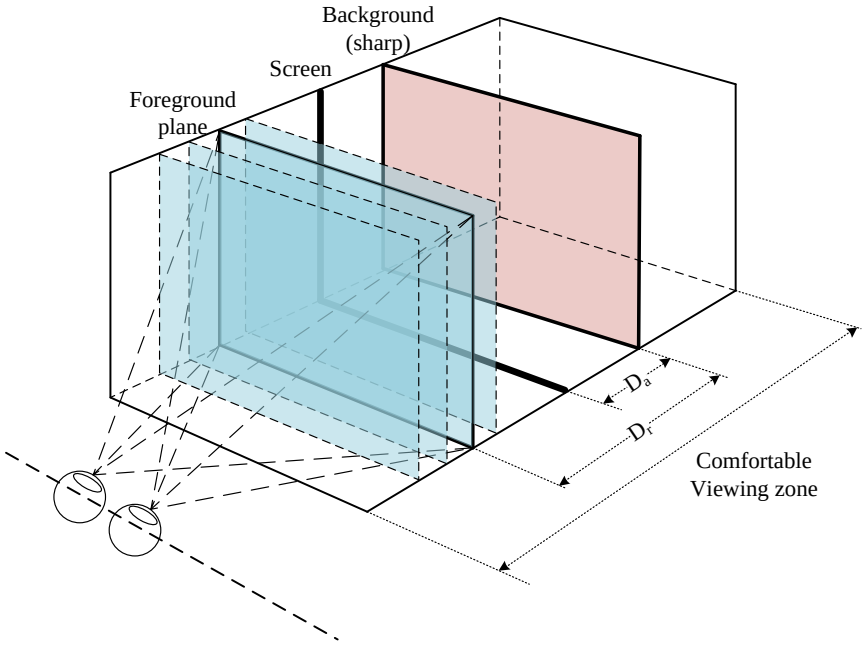
Once the absolute position and the relative distance were selected for each BB-stimulus, then we paired the BB-stimulus with a set of 7 or 8 SB stimuli which are with relative distances ranging from 4 cm less than the BB stimulus to 8 cm larger than the reference stimulus. These steps were chosen considering both the depth rendering ability of the screen and the depth perception ability of the observers.

The experiment consists of 155 trials in total. Each trial contains a BB-stimulus and a SB-stimulus. This setup is shown in Figure 9.3.2.

The subjective experiment was conducted using a two-alternative forced choice task. For each trial, one of the 155 conditions was chosen randomly and a pair of stimuli were displayed. Observers were asked to look at the butterfly and then determine in the trial whether the BB-stimulus contained a larger depth interval between the butterfly and the background than the SB-stimulus. As a two monitor screen setup was technically



(a)



(b)

Figure 9.3.2: Schematic diagram of the experiment setup. The blue planes represent the (foreground) depth planes in which the butterfly is located. The red plane represents the depth plane in which the background is located. We named the distance between background plane and the screen as Absolute Position (D_a), while the interval between the foreground plane and the background plane as the Relative Distance (D_r). The variation range of these two planes always stayed in the comfortable viewing zone. (a) shows the BB-stimulus case; (b) shows the SB-stimulus case. The selection of the parameters D_a and D_r for both BB-stimuli and SB-stimuli are presented in Table 9.1.

D_a (cm)	$D_{r_{BB}}$ (cm)	$D_{r_{SB}}$ (cm)							
-19.7	0		0	1.2	2.3	3.4	4.5	5.6	6.6
-19.7	6.6	3.4	6.6	7.6	8.6	9.6	10.6	11.5	13.4
-19.7	13.2	11.1	13.2	14.3	15.6	16.9	18.1	19.7	21.7
-19.7	19.7	17.3	19.7	20.9	22	23.5	24.9	26.6	28.3
-19.7	26.3	24.9	26.3	27	28	29.3	30.5	32	33.4
-19.7	32.8	30.5	32.8	34	35.1	36.2	37.5	38.7	39.9
-13.2	0		0	1	2	3	4	4.9	5.9
-13.2	6.6	4.5	6.6	7.7	8.6	9.8	11.1	12.7	14.3
-13.2	13.2	11.5	13.2	14.3	15.4	16.5	17.6	18.7	19.7
-13.2	19.7	18.3	19.7	20.7	21.4	22.3	23.3	24.5	26
-13.2	26.3	24.5	26.3	27.1	27.9	29	30.1	31.4	15.4
-6.6	0		0	0.9	1.8	2.7	3.5	4.4	5.2
-6.6	6.6	3.5	6.6	8	9.1	10.2	11.6	13.4	15.4
-6.6	13.2	12	13.2	14.4	15.4	16.3	17.3	18.5	19.6
-6.6	19.7	16.9	19.7	20.8	21.6	22.7	23.8	25.3	26.8
0	0		0	0.8	1.6	2.3	3.4	4.5	5.6
0	6.6	4.5	6.6	7.6	8.6	9.5	10.8	12.3	14.3
0	13.2	11.4	13.2	13.7	14.3	15.1	15.9	17.2	18.7
6.6	0		0	1	2	2.9	3.9	4.8	6.3
6.6	6.6	4.2	6.6	7.1	8	8.8	9.9	11.2	12.9

Table 9.1: The selection of parameters D_a and D_r for both BB-stimuli and SB-stimuli. Note that there are only 7 possible selections with the $D_{r_{BB}} = 0\,cm$ cases, because putting the foreground behind the background will cause trouble of fusion.

not possible, observers controlled the displaying of stimuli by means of a key press to switch from one stimulus to the other. When the observers switched between the stimuli, a 700 ms interval of grey color background was shown in order to avoid memorization by the observers of the "exact" positions of the foregrounds. For each trial, the total observation time and the number of switches were not limited.

9.4 Result and Analysis

For each condition, some observers considered the BB-stimulus as having a larger depth interval (between the foreground and the background), while the other observers chose the SB-stimulus. We measure the proportion of "*BB-stimulus contains a larger depth interval*" responses. The result is plotted as a function of the disparity difference between $D_{r_{BB}}$ and $D_{r_{SB}}$, which represents the D_r in the BB-stimulus and the D_r in the SB-stimulus, respectively.

Cumulative Weibull function is used as the psychometric function. The disparity difference corresponding to the 50% point is considered as the Point of Subjective Equality (PSE). When measuring the disparity difference at that point, the increase of perceived depth can be obtained. In total, by filtering out the data of 7 observers who made decisions in the test quite different from other observers, 28 observations of each condition were included in the computation. An example pattern of response and the fitted psychometric function is shown in Figure 9.4.1.

According to the setup of the experiment, blur was added to the backgrounds being located at various absolute depth positions in different trials, while the depth interval between the background and the foreground object stayed also as a free parameter. We first plot the curve of the PSE as a function of the background disparity (D_a) which represents the depth of the blurred background (see Figure 9.4.2). Each of the five points on the curve is obtained by considering all the possible depth intervals ($D_{r_{BB}}$) which are with the same background depth (D_a). The steps are slightly different because the same distances equal to various values of disparity in angular units depending on the viewing distance.

As we can see in the Figure 9.4.2, the PSE curve shows a clear offset from the unit slope. This offset indicates an increase of perceived depth caused by blur in binocular vision. Note that the increase of perceived depth is almost constant (approximately 180 *arcsec*) regardless of the variation of absolute depth of the background. This phenomenon means that the increase of perceived depth caused by blur is insensitive to the disparity of the blurred background, e.g. the absolute position. This phenomenon might be due to the fact that all the blur actually exists only on the screen plane where human eyes accommodate.

We plot also the curve of the PSE as a function of the depth interval (in length unit) between the foreground object and the background (see Figure 9.4.3a). In Figure 9.4.3a, each of the five points on the curve is obtained by considering all the possible values

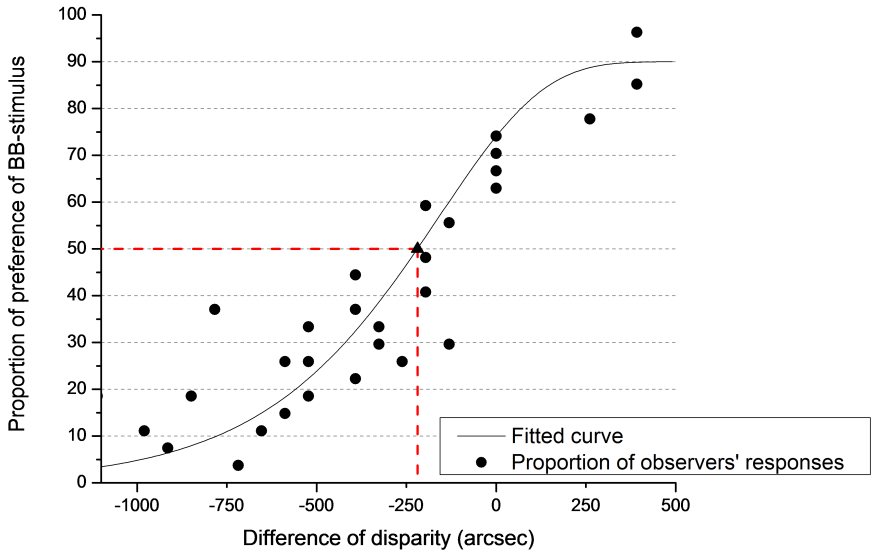


Figure 9.4.1: An example pattern of the proportion of observers' responses and the fitted psychometric function. In this trial, we consider $D_{r_{BB}}$= 6.6 cm and D_a= -19.7 cm, -13.2 cm, -6.6 cm, 0 cm, 6.6 cm. An equal apparent depth is reached at -220 *arcsec*, which corresponds to the second point in Figure 9.4.3b.

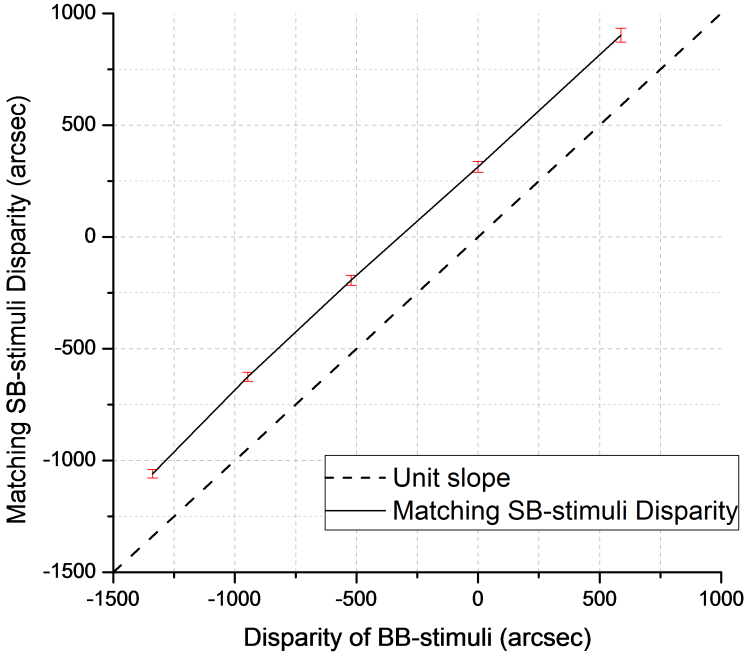
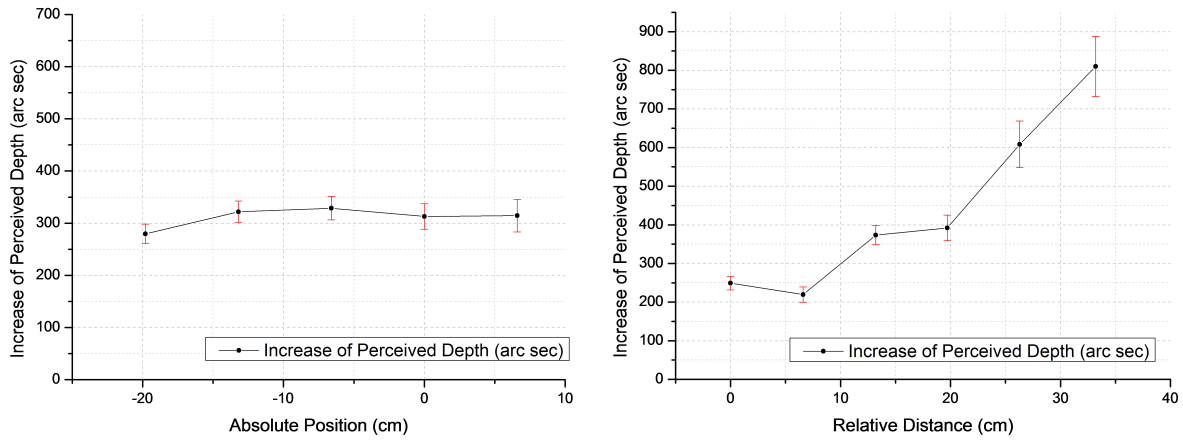


Figure 9.4.2: The PSE curve shows subjective matches between SB-stimuli and BB-stimuli. Negative values denote far disparities. The dashed line at unit slope indicates physically correct disparity matches.

of background depth (D_a) for one depth interval ($D_{r_{BB}}$). The value of 0 cm means that the foreground and the background are at the same depth plane. There exists no difference of disparity between them. Note that according to the viewing geometry, the same depth intervals stand for slightly different disparities, therefore we plot the increase of perceived depth as a function of depth interval which is represented in length unit.

The curve in Figure 9.4.3b illustrates a clear increase of perceived depth which is created by the influence of blur. We can also find that the increase of perceived depth varies positively with increasing depth interval between the fixated sharp foreground object and the defocused blurred background.

When the foreground object is close to the background (the relative distance is 0 cm or 6.6 cm), the influence of blur is relatively small. The 0 cm depth interval comes with a slightly larger increase of perceived depth than the 6 cm depth interval. The reason might be because when the depth interval is 0 cm, it seems to the observers that they are looking at a planar image with a sharp foreground and blurred background. For this kind of image, intentional blur has been introduced by photographers for a long time to induce some illusions of the existence of depth between the foreground and the background. When disparity differences appear in the image, observers start to notice that they are facing a stereoscopic image. This might be the reason why the 0 cm depth interval shows a slightly larger increase of perceived depth than the 6 cm depth interval. Starting from the second point (6.6 cm depth interval), the enhancement is monotonic.



(a) (b)

Figure 9.4.3: Results from the experiment. The left figure shows the increase of perceived depth as a function of absolute position, while the right figure shows the increase of perceived depth as a function of relative distance. Note that each point on the curves might be obtained from different number of observations.

As shown by the curves in Figure 9.4.3, the overall tendency of how the depth interval ($D_{r_{BB}}$) and the absolute position of the background (D_a) affect the perceived depth

seems easy to understand. However, we split the data from Table 1 and computed the PSEs of every combination of $D_{r_{BB}}$ and D_a, we then found that the variations of PSEs are with some uncertainties. In Figure 9.4.4, we plot a surface (consisting of blue points) which is obtained from the curve in Figure 9.4.3b, while the red points come from the PSEs which are computed individually by each combination of $D_{r_{BB}}$ and D_a. The figure shows that the changes of perceived depth do not vary strictly according to the tendency we described previously. There exists a lot of variability at certain points. Namely, despite the overall tendency of how the blur affects perceived depth at different relative distances and absolute positions is known, the predictions of perceived depth of objects at certain positions or the perceived depth of certain observers remain uncertain. Several types of analysis were performed in order to learn about the source of these aberrations. However, there was no indication found concerning a bi- or multimodal distribution of the observers. So the variance that is found in our experiment concerning the individual data points has to be considered as measurement noise for the moment.

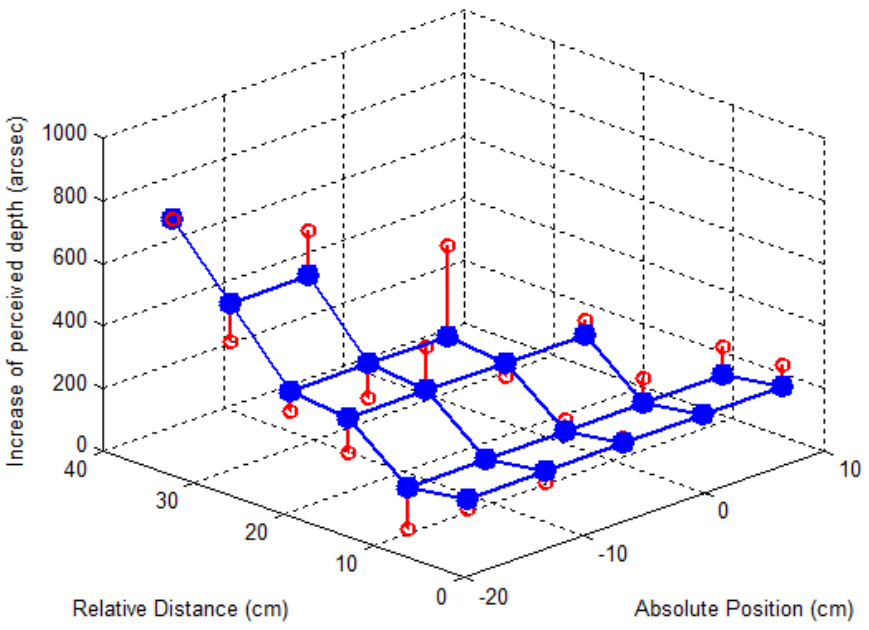


Figure 9.4.4: The surface consisting of blue points is generated from the curve in Figure 9.4.3b. We ignore the variance in the curve in Figure 9.4.3a, because the variance is relatively small. The red points are obtained from the PSEs which are computed individually by each combination of relative distance ($D_{r_{BB}}$) and absolute position (D_a).

9.5 Discussion

Remind that one purpose of the present study is to try to increase the Quality of Experience (QoE) of 3DTV. However, how many factors actually influence the QoE of 3DTV, and the weight of each of these factors remain open questions. For instance, viewing geometry, crosstalk, and distortion can all be some important factor affecting the QoE. Nevertheless, it would be reasonable to consider that the QoE of 3DTV might contain at least three dimensions: the visual comfort, the depth perception, and visual quality.

Based on the 3D images processing approach proposed in the present study, we can decrease the binocular disparity to reduce the conflict between accommodation and vergence, so the visual comfort is supposed to be increased. On the other hand, blur is introduced to compensate the loss of apparent depth, so the depth perception can be supposed to be kept. The remained question concerns the variation of visual quality.

Due to the additional blur, the content of image has been changed. However, this change of content does not necessarily lead to a decrease of visual quality. Whether the overall visual quality of the image is affected (or decreased) might depend on whether the blurred part of the image contains regions of interest. If the regions of interest are kept sharp, the influence on visual quality can be minor. But if any region of interest is blurred out, important information of the scene is thus lost. A considerable influence of visual quality is thus introduced.

Another problem to be solved is whether the creation of blur changes the locations of regions of interest. It has been demonstrated that there exists a strong relationship between blur and deployment of visual attention in 2D viewing condition [Khan 11]. The creation of blur may generate or eliminate regions of interest. Therefore, a correct detection of region of interest before introducing blur may help to solve this problem. The development of a reliable computational model of 3D visual attention becomes thus a critical issue.

9.6 Conclusion

The influence of defocus blur on the perceived depth in a stereoscopic scene and its relationship with binocular disparity were studied. The experimental result indicates that image blur contributes to increasing perceived depth in stereoscopic images, when measured against perceived depth in stereoscopic images without the additional blur. The increase of depth can be considered as a function of the relative distance between the fixated foreground object and the blurred background, while this increase is insensitive to the distance between the viewer and the depth plane at which the blur is added.

The feasibility of enhancing the perceived depth by reinforcing a monocular cue, namely defocus blur, provides an interesting way to deal with the conflict between accommodation and vergence when 3D images are shown on a planar stereoscopic display. Generally, the foreground object popping out of the screen is the most important object

in the scene, while the large disparity of the object may lead to visual discomfort when it is actually fixated by the observer. Our results show that it is possible to decrease the disparity of this object without losing its pop-out effect by adding some blur on its background.

Key points

Context

- ❏ Stereoscopic 3D displays nowadays enhance only one cue, the binocular disparity, to increase the sensation of depth. The conflict between accommodation and vergence is thus elicited. This conflict is generally considered as the main reason of visual fatigue and visual discomfort when people are watching stereoscopic 3D content.
- ❏ Different depth cues work with each other in an adaptive way. The perceived depth can be affected by adjusting one or several depth cues independently.

Contributions

- ❏ We propose an idea that one can decrease the binocular disparity to limit the visual discomfort in 3DTV, and compensate the loss of perceived depth by increased a monocular depth cue, defocus blur.
- ❏ We conducted a psychophysical experiment to quantify how the combination of blur and disparity affects the perceived depth. Compared to the previous studies in the literature, we use a state-of-the-art stereoscopic display system and more natural stimuli.
- ❏ Our experimental result indicates that image blur makes contributions to the impression of depth perceived in stereoscopic images. The increase of depth can be considered as a function of the relative distance between the fixated foreground ground object and the blurred background.

Chapter 10

Conclusion and perspectives

10.1 Summary and Contribution

In this thesis, we present studies focusing on several aspects of the research of visual attention. The framework of the thesis can be mainly divided into two parts. The first part of this thesis concerns the issues related to ground truths of visual attention. In the second part, our studies are related to the modeling of visual attention for 3D viewing condition. A summary of our work and contribution is presented as follow.

Verify the reliability of ground truth

Our work starts with identifying the reliability of FDM from different eye-tracking databases. In order to evaluate the similarity of FDM created from independent experiments, we start an international collaboration, which enables us to compared the FDM obtained from three experiments conducted in thee different countries. We particularly focused on the influence of visual content and image presentation time. The reliability of the FDM as a ground truth has also been validated on three image processing applications: visual saliency modeling, image quality assessment and image retargeting. Based on the image database we used, we found that the FDM from the three laboratories were very similar. The impacts on the three applications that we study is low. These findings suggest that FDM from independent eye-tracking experiments can indeed be considered as reliable ground truths for these three image processing applications.

Quantify the relationship between two types of ground truth

Our second study is to quantitatively identify the similarities and difference between visual saliency maps (i.e. fixation density maps) and importance maps, which are two widely used ground truth for attention-related applications. We start our work from the perspective concerning the two attention mechanisms: bottom-up and top-down.

We thus conduct two psychophysical experiments. In the first experiment, importance maps and the category of each object (main subject, secondary object, or background) were collected by asking human subjects to rate the importance value of each object within hand-segmented images. In the second experiment, fixation density maps were obtained by a task-free eye-tracking experiment. By comparing the importance maps with the saliency maps, we found that the two types of maps are related, but perhaps less than one might expect. The saliency maps were shown to be effective at predicting the main subjects. However, the saliency maps were less effective at predicting the objects of secondary importance and the unimportant objects. We also found that the vast majority of early gaze position samples (0-2000 ms) were made on the main subjects, suggesting that a possible strategy of early visual coding might be to quickly locate the main subject(s) in the scene.

Create a new eye-tracking database for 3D content

To solve the problem of lacking ground truth in the community of 3D visual attention modeling, we conduct a binocular eye-tracking experiment. We create a new eye-tracking database containing eighteen 3D natural-content images (as well as their corresponding 2D version), the corresponding disparity maps, and eye movement data (including fixation density maps for both 2D and 3D version) of both eyes. This database helps in solving the problem of lacking ground truth in the research area of 3D visual attention modeling.

Influence of depth on 3D visual attention: a study of depth-bias

Comparing the viewing of 3D content with the viewing of 2D content, the change of depth perception largely changes human viewing behavior. The goal of our study is to determine if there exists a so-called "depth-bias" in the viewing of 3D content on a planar stereoscopic display. We conducted an eye-tracking experiment using a state-of-the-art stereoscopic display and eye-tracker. A large number of synthetic stimuli were designed for the experiment in order to get rid of the effect of 2D visual features, and let the visual attention of observers be influenced by only depth information. We examined how the depth order and the relative depth of the objects influenced observers' viewing behavior. Experimental results clearly show that observers payed more and earlier attention to the objects closest to them than to the other objects. Results appear to demonstrate the existence of the depth-bias in the viewing of 3D content on a planar stereoscopic screen.

A depth-feature-based model of 3D visual attention

We first do a bibliography study which introduces state-of-the-art computational models of 3D visual attention in the literature, and propose a new taxonomy of these 3D attention models according to the way of using depth information. We then propose a

new depth-saliency-based model of 3D visual attention. Bayes's theorem is applied for learning from previous eye-tracking data in order to compute the depth saliency map. The results demonstrate a large added value of the depth saliency map and a good performance of the proposed depth-saliency model. Two different ways of applying depth information in a 3D visual attention model are compared in our study. Our results show that, creating a depth saliency map based on depth feature achieves a higher performance than a simple depth-weighting method (a multiplication of 2D saliency map and depth map). This result indicates the importance of the depth feature extraction in modeling 3D visual attention.

Integration of center-bias into 3D attention modeling

We study the variation of center-bias from 2D viewing to 3D viewing, and the way of integrating the center-bias into a 3D visual attention model. Our work demonstrates that center-bias in 3D viewing condition is slightly weaker than 2D viewing condition. We propose a simple 3D attention model which (1) exploits existing 2D visual attention model by an "attention-shifting" process and (2) integrates center-bias which is tuned for 3D viewing condition. Our results demonstrate a large added value of center-bias in predicting saliency map of 3D images. We also propose a hybrid model integrating 2D saliency, depth saliency and center-bias. Both qualitative and quantitative assessments demonstrate a good performance of the proposed hybrid model on 3D images.

Towards a potential application of 3D visual attention

Visual attention is assumed to be valuable in improving the Quality of Experience of 3D-TV when collaborating with blur. However, the impact of depth cue integration of disparity and blur still remains as open question. We conducted a psychophysical experiment to study the influence of defocus blur on depth perception in a stereoscopic scene and its relationship with binocular disparity. Our results demonstrate that image blur makes contributions for increasing perceived depth in stereoscopic images, when measured against perceived depth in stereoscopic images without the additional blur. The increase of perceived depth can be considered as a function of the relative distance between the fixated foreground object and the blurred background, while this increase is insensitive to the distance between the viewer and the depth plane at which the blur is added.

10.2 Limitations and perspectives

A large part of our work presented in this thesis is to study visual attention by means of psychophysical experiments. During our research, we found out that there were still

limitations in our work; some issues still remain as open questions and need to be studied in the future work:

- With respect to the verification of FDM's reliability, it is still difficult to quantify the degree to which each factor is affecting the FDM. It is therefore instrumental to extend this work by conducting experiments conjointly with careful variation of certain factors. It further needs to be verified whether a larger number of participants would result in even more stable FDM and thus in higher similarity between the experiments. Thresholds need to be determined that specify the minimum number of participants for FDM with given similarity constraints.

- There are still ways to improve the algorithms that predict the category of objects based on saliency map. The timing data suggests that a multi-stage predictor might be more effective than our single-stage attempts. In the first stage, the gaze position samples from 0-2000 ms (or another early time slice) would be used to determine the most important objects (main subjects). This knowledge of the main objects would then be used to guide subsequent predictions for other objects.

- In terms of the ground truth in 3D visual attention modeling, an eye-tracking database containing only eighteen images is still not enough. An eye-tracking database containing a larger number of images, or 3D videos, needs to be created in the future. On the other hand, new technique of calibration, e.g. the "volumetric calibration" [Huynh-Thu 11a], is also necessary to be introduced in eye-tracking experiments using 3D stimuli.

- In the study of investigating depth-bias, we only focus on the influence of relative depth on the distribution of visual attention. The study of the influence of absolute depth (e.g. disparity) is not included yet. However, absolute depth might be another important factor affecting the depth-bias, since the existence of disparity-selective neurons in the primary visual cortex (V1) has been demonstrated [Neri 99, Barlow 67, Nikara 68, Poggio 77]. A perspective of the presented study is to take into account absolute depth information.

Another main contribution of the thesis is a new depth-saliency-based model of 3D visual attention. However, the proposed model still has several limitations. For instance, the proposed model takes into account only one depth feature (i.e., depth contrast); only a simple pooling strategy is used to combine the depth saliency map and 2D saliency map. In the future work, one might consider to include more depth features (e.g. surface curvature or depth gradient). A more sophisticated pooling strategy might also improve the performance of the model.

Appendix

Appendix A

Publications

Journal

1. **Junle Wang**, Matthieu Perreira Da Silva, Patrick Le Callet, Vincent Ricordel, "A computational model of stereoscopic 3D visual attention", *IEEE Transaction on Image Processing.* (In second round review)

2. Ulrich Engelke, Hantao Liu, **Junle Wang**, Patrick Le Callet, Ingrid Heynderickx, Hans-Jürgen Zepernick, "A comparative Study of Fixation Density Maps", *IEEE Transaction on Image Processing*, (Accepted).

3. **Junle Wang**, Patrick Le Callet, Sylvain Tourancheau, Vincent Ricordel, Matthieu Perreira Da Silva, "Study of depth bias of observers in free viewing of still stereoscopic synthetic stimuli", *Journal of Eye Movement Research*, 5(5):1, 1-11.

4. Hantao Liu, Ulrich Engelke, **Junle Wang**, Patrick Le Callet, Ingrid Heynderickx, "Variation in Visual Attention between Observers and its Impact on Objective Image Quality Assessment", *IEEE Transactions on Circuits and Systems for Video Technology*, (submitted).

Conference

1. **Junle Wang**, Matthieu Perreira Da Silva, Patrick Le Callet, Vincent Ricordel, "Study of center-bias in the viewing of stereoscopic image and a framework for extending 2D visual attention models to 3D", Human Vision and Electronic Imaging XVIII, San Francisco, USA, 2013 (Accepted).

2. **Junle Wang**, Patrick Le Callet, Vincent Ricordel, Sylvain Tourancheau, "Quantifying depth bias in free viewing of still stereoscopic synthetic stimuli", *16th European Conference on Eye Movements*, Marseille, France, 2011.

3. **Junle Wang**, Marcus Barkowsky, Vincent Ricordel, Patrick Le Callet, "Quantifying how the combination of blur and disparity affects the perceived depth", *Human Vision and Electronic Imaging XVI, Proceedings of the SPIE*, Volume 7865, pp. 78650K, San Francisco, USA, 2011.

4. **Junle Wang**, Marcus Barkowsky, Vincent Ricordel, Patrick Le Callet, "Quantifying how the combination of blur and disparity affects the perceived depth", *Sino-French Workshop on Research Collaborations in Information and Communication Technologies (SIFWICT*), Nantes, France, 2011.

5. **Junle Wang**, Damon M. Chandler, Patrick Le Callet, "Quantifying the relationship between visual salience and visual importance", *Human Vision and Electronic Imaging XV, Proceedings of the SPIE*, Volume 7527, pp. 75270K , San José, USA, 2010.

6. Hantao Liu, **Junle Wang**, Judith Redi, P. Le Callet, I. Heynderickx, "An efficient No-Reference metric for perceived blur", *3rd European Workshop on Visual Information Processing*, Paris, France, 2011.

7. Jing Li, Marcus Barkowsky, **Junle Wang**, Patrick, Le Callet, "Study on visual discomfort induced by stimulus movement at fixed depth on stereoscopic displays using shutter glasses", *17th International Conference on Digital Signal Processing*, Corfu, Greece, 2011.

Project report

1. Vincent Ricordel, **Junle Wang**, Josselin Gautier, Le Meur Olivier, Emilie Bosc, "Perceptual modelling for 2D and 3D", Projet ANR PERSEE

2. Patrick Le Callet, Vincent Ricordel, **Junle Wang**, Josselin Gautier, Christine Guillemot, Laurent Guillo, Olivier Le Meur , Emilie Bosc, Luce Morin, Marco Cagnazzo, Beatrice Pesquet-Popescu, "2D/3D Codec architecture" , Projet ANR PERSEE.

Appendix B

Depth perception in stereoscopic 3D content

Nowadays, stereoscopic 3D content increases the sensation of presence through the enhancement of depth perception. To achieve this task, binocular depth cues, such as binocular disparity, are introduced and fused together with other monocular depth cues in an adaptive way depending on the viewing space conditions. Studies related to stereoscopic 3D have been recently gaining an increasing amount of attention because of the emergence of 3D content (in cinema and home) and recent availability of high-definition 3D-capable acquisition and display equipments.

From this chapter, the study starts to focus on the topics related to the deployment of visual attention in three-dimensional viewing condition. Before going into the computational modeling of stereoscopic 3D visual attention, studies related to depth perception in stereoscopic 3D viewing condition is firstly introduced in this section.

B.1 Depth perception

To perceived the three-dimensional real world, a vast amount of information is processed by the brain. Such information is believed to comprise the so-called "depth cues". Visual system uses these depth cues to estimate the depth information of a scene. The importance and the priority of each cue vary depending on the viewing condition. For instance, the available depth cues in a scene are given higher priority by the visual system when other depth cues are missing. On the other hand, the influence of one depth cue may largely vary according to the viewing distance from viewer to the object.

The inconsistency of importance and priority of depth cues makes the classification and ranking of depth cues difficult. There exists different ways to classify depth cues [McAllister 93]. Nevertheless, according to the number of eyes involved, the depth cues can be divided into two categories (see Table B.1, [Mather 09]): (1) monocular cues, which are available when only one eye is used, and are also available when both eyes are

used; (2) binocular cues, which are available only when both eyes are used together.

Monocular cues	**Binocular cues**
Retinal image size	Vergence
Height in the visual field	Binocular disparity
Texture gradient	
Defocus blur	
Atmospheric perspective	
Accommodation	
Motion parallax	
Shadows and shading	
Occlusion and Interposition	

Table B.1: List of depth cues.

Monocular cues

- Retinal image size. It arises from the geometric fact that an object of fixed actual size will project a progressively smaller image onto the retina as it is viewed from more distant viewpoints. This depth cue requires knowledge of the object's actual size (see Figure B.1).

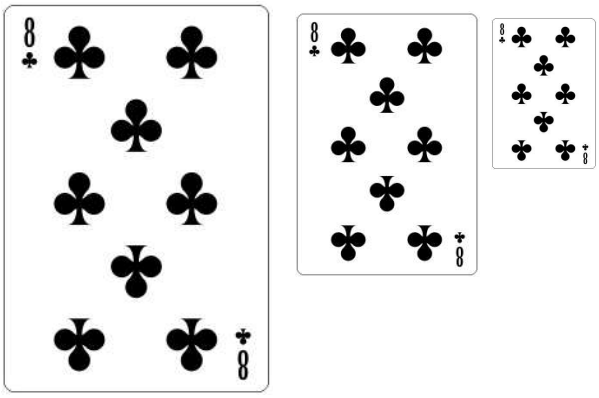


Figure B.1: Example of the effect of retinal image on perceived depth.

- Height in the visual field. This is a depth cue based on the vertical position of a point in the visual field. Higher positions correspond to further distance. For instance, if the observer is standing on a level surface outdoors (a so-called "ground plane") and looking toward the distant horizon, a higher position in the visual field of an object resting on the surface means that the object is further located.

- Texture gradient. This is a depth cue to the orientation and the depth of a textured surface. It is based on graded variation in the size, shape, and density of texture elements (see Figure B.2).

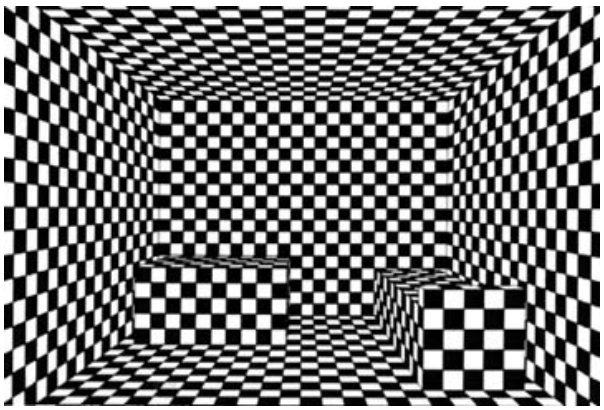


Figure B.2: Example of the effect of texture gradient on perceived depth.

- Defocus blur. Like cameras, human eye has a limited depth of field. If the eye is focused on a point at a given distance, in the retinal image, some parts of the scene will inevitably be more spatially blurred than others. Specifically, outside a certain range, points nearer or further away than the point of focus can not be seen sharply. As a consequence, blur of the image offers a cue to the distance. Defocus blur is a quantitative depth cue, the relationship between defocus blur and perceived depth is introduced in detail in Section 9.2.

- Atmospheric perspective. This depth cue is based on the fact that light from distant objects has to travel through more atmosphere than light from nearby objects. Therefore, distant objects appear reduced in contrast, and a slight shift toward bluish hues (see Figure B.3).

- Accommodation. This nonvisual cue concerns the change of the lens of the eye. It is controlled by the ciliary muscles to maintain a sharply focused image of the fixated point. Fixation on a relatively near point corresponds to a relatively released state of the muscles. Therefore, information on the state of the ciliary muscles provides the information of absolute fixation distance.

- Motion parallax. This depth cue exists in a dynamic scene. It concerns the movement in one part of the scene relative to another. This relative movement is produced by objects moving at different distances from the viewpoint.

- Shadows and shading. Directional illumination creates shadows. Due to the fact that visual system tends to assume that light is directed from above, shadows and shading can provide depth information. An example is shown in Figure B.4.

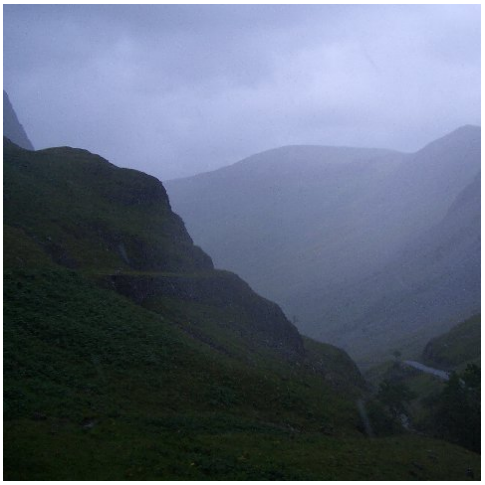


Figure B.3: Example of the effect of atmospheric perspective on perceived depth.

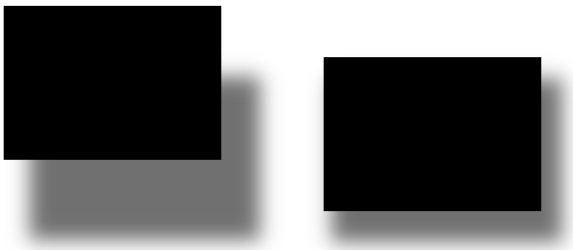


Figure B.4: Example of the effect of shadows and shading on perceived depth.

- Occlusion and interposition. This is an ordinal depth cue, which is based on partial occlusion of a far object by a near object. People prefer to consider the object with a more continuous edge as a closer object (see Figure B.5).

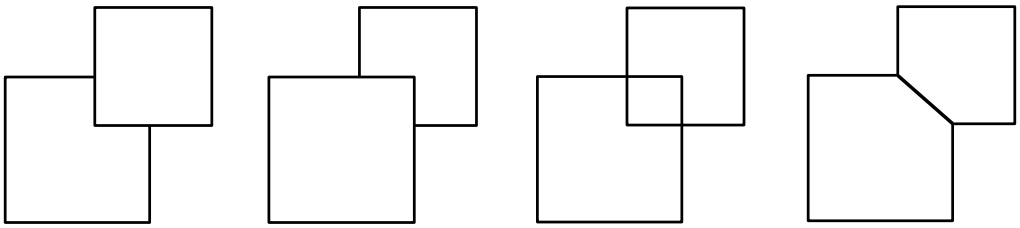


Figure B.5: Example of the effect of occlusion and interposition on perceived depth. In the first and the second image, the ordinal depth of the two squares can be easily perceived. In the third and the forth image, it is difficult to identify the object's depth.

Binocular cues

- Vergence. This is a nonvisual cue. When both eyes fixate on the same point in space, the visual axes of the two eyes converge by a certain angle. This vergence angle form by the intersection of the two visual axes offers a cue to fixation distance.

- Binocular disparity. Humans eyes are typically 6.3 cm apart. Consequently, the retinal images both eye are slightly different from each other. These differences cause a powerful depth cue, binocular disparity, which leads to stereo vision. Binocular disparity is a quantitative depth cue, the relationship between binocular disparity and perceived depth is introduced in detail in Section 9.2.

B.2 Conflicts in stereo vision

Among a set of depth cues, stereoscopic 3D displays enhance particularly the binocular disparity to create stereo vision for the viewers. However, people claim that watching stereoscopic 3D content causes a higher level of visual fatigue and visual discomfort compared to the viewing of 2D content. Clearly, if the stereoscopic 3D technique is to succeed in commercial market, this problem of visual fatigue and visual discomfort needs to be solved.

It is believed that the problem of visual fatigue and visual discomfort is caused by the imperfect simulation of depth cues on planar stereoscopic 3D display. In real world, a scene contains various depth cues that specify the same 3D layout. However, the stereoscopic 3D displays, which try to present a three-dimensional scene, are flat. Images can be only displayed on one surface. As a consequence, the depth cues originally

contained in the real-world are introduced differently by the display. Some depth cues are created by the stimuli themselves, while some other depth cues are created by the display. This difference of the sources of depth cues generates conflicts.

One important conflict among depth cues is the conflict between accommodation and vergence [Hoffman 08]. This conflict is generally considered as the main reason of visual fatigue and visual discomfort when watching 3D content on a stereoscopic 3D display.

Both accommodation and vergence are nonvisual depth cues. As introduced in Section B.1, accommodation is controlled by the ciliary muscles which adjust the focal distance of the eyes by changing the shape of the lens. To keep the retinal image sharp, the focal distance of accommodation should be kept within in a certain range around the object. This range is called depth of field, which has a value of 0.3 diopter under normal circumstances. On the other hand, to keep the stimuli fused, the eyes should converge to a distance close to the object distance, within a tolerance range. This range is called Panum's fusion area, which has a value of 15-30 arcmin.

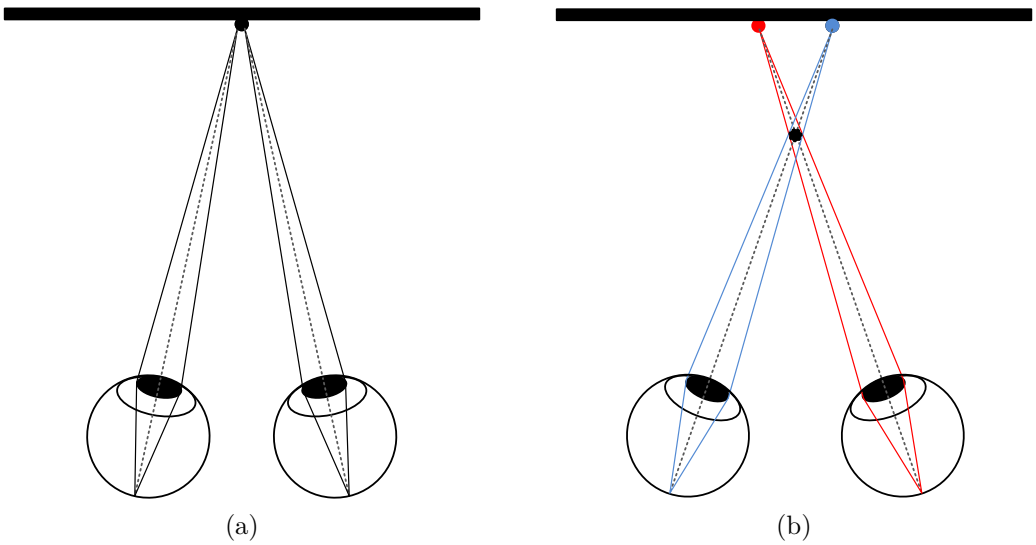


Figure B.1: Illustration of the accommodation-vergence conflicts.

In the real world, accommodation and vergence are normally coupled. Accommodative changes can evoke vergence changes, and vergence changes can also evoke accommodative changes [Fincham 57]. Focal and vergence distances are always close. When watching 2D content on a planar display, the situation is similar (see Figure B.1a). The two eyes accommodate and converge on the same location (i.e. the stimuli on the screen plane). No conflict is introduced.

When watching 3D displays, accommodation and vergence may be decoupled. As illustrated in Figure B.1b, focal distance becomes different from vergence distance. The viewer's vergence point varies depending on the stimuli while the viewer has to always accommodate at the screen plane to keep the stimuli being sharply focused. When the

difference between focal distance and vergence distance is over the tolerance range, the relationship between accommodation and vergence is disrupted, and creates the conflict.

For the reasons stated above, the disparity of stereoscopic content needs to be within a certain range to limited the accommodation-vergence conflict. This extend of this range, which is called "comfortable viewing zone", is computed based on the viewing condition [Chen 10].

Appendix C

Image quality assessment and visual attention

C.1 Introduction

When an image is supposed to have been processed by image enhancement or lossy compression, for the purpose of storing or transmission, it is necessary to measure the quality of the distorted image. The measurement of image quality can be achieved in two different ways, namely subjective quality assessment and objective quality assessment. Subjective quality assessment regards to scoring experiments conducted for the purposed of evaluating image quality. Due to the participation of human observers, the subjective quality assessment is precise but cost. On the other hand, the objective quality assessment consists of the use of computational model called "objective quality metric" (OQM), which aims to automatically produce quality scores close to subjective quality scores given by human observers during subjective quality assessment test. The objective metrics are designed to serve eventually as an alternative for expensive quality evaluation by human observers.

In the literature, a body of OQM have been investigated. They range from specific quality metrics that focus on a particular type of distortion, to general metric that aims to assess the overall perceived quality. Most of the OQM can be classified into three categories: full-reference (FR) metrics, reduced-reference (RR) metrics, and no-reference (NR) metrics, depending on to what extent they use the original image. Advances in image quality assessment have shown the need and practical attainability of integrating relevant aspects of the human visual system (HVS) in the design of OQM. In the literature, lower level aspects of the HVS, such as contrast sensitivity and masking were successfully modeled and included in various objective metrics [Wang 06, Hantao 11]. Researchers now attempt to further improve the prediction performance of OQM by taking into account higher level aspects of the HVS, such as visual attention [Le Meur 10b, Engelke 11, Moorthy 09].

The mechanism and the influence of visual attention for image quality assessment have not been fully understood yet. Several studies (e.g. [Liu 11, Le Meur 10b, Engelke 09a]) have been proposed in the literature, regarding to the incorporation of visual attention into various objective metrics. Most of them incorporate visual attention in an *ad hoc* way, based on optimizing the performance increase in predicting perceived quality [Liu 11]. These studies assume that a distortion occurring in an area that attracts more observer's attention is more annoying than the distortion in other areas. They focus on the extension of an OQM with a (so-called) computational attention model; a map representing the spatial distribution of image distortions is weighted with the calculated saliency (see, e.g. in [Moorthy 09]). This process is referred to as "visual importance pooling", and has been proven to result in a performance gain in OQM.

Nevertheless, in such an approach, the evaluation of the added value of saliency may heavily depend on the accuracy and reliability of the attention model used. To better understand the intrinsic added value of including visual attention in OQM, researchers used actual visual attention data obtained from eye-tracking recordings (see, e.g. in [Le Meur 10b]). Ninassi et al. [Ninassi 07] incorporates eye-movement data recorded during a quality assessment task into two full-reference OQM (i.e. the peak signal-to-noise ratio and structural similarity), and no clear improvement is found. On the other hand, Liu et al. [Liu 11] focus on the added value of natural scene saliency (NSS), which means the saliency driven by the original image content. The NSS can be obtained by means of eye-tracking experiments during free-viewing task. They demonstrated that the addition of NSS was beneficial to prediction performance of OQM. The results, however, also showed that the actual amount of gain in prediction accuracy depended on several factors, among which were the saliency map used, image content, the distortion type and the objective metric itself.

In this study, we focus on how the image content dependency of NSS affects the added value of NSS in objective image quality assessment. Knowledge on this issue may be highly beneficial for the development of a more reliable image quality assessment strategy adaptively incorporating saliency into objective metrics, depending on the visual content. It has also been demonstrated that the variation in NSS between participants largely depends on the visual content [Liu 11].

Our study is based on three similar eye-tracking experiments independently conducted in different laboratories[1]. As the image used in all the three eye tracking experiments are taken from the LIVE image quality database, we have as well a large set of distorted images and their respective mean opinion scores (MOS) available for the design and validation of the quality models. We first analyze to what degree the improvement of three quality prediction models, the Peak Signal-to-Noise Ratio (PSNR) [Wang 06], Structural Similarity (SSIM) Index [Wang 04], and Visual Information Fidelity (VIF)

[1]This study is performed though an international collaboration with Philips Research Laboratories (The Netherlands), Delft University of Technology (The Netherlands), and University of Western Sydney (Australia).

criterion [Sheikh 06], varies with the FDM of different databases used. Secondly, we investigate to what extent the variation in NSS between individual human observers is a reliable measure to classify image content. Thirdly, we evaluate whether the extent of such variation indeed affects the actual gain in prediction accuracy that can be obtained by including saliency in objective metrics.

C.2 Experiment

The eye-tracking data in previous studies (e.g. [Liu 11]) showed also that the variation in NSS between observers strongly depended on the image content. To verify the generalization of these relations, we evaluated the similarity in eye-tracking data for three experiments that were conducted independently in different laboratories (see Chapter 3). With the inter-laboratory comparison we wanted to provide solid evidence on (1) to which extend the FDM obtained in different experiments affect the performance of OQM that have integrated NSS; and (2) whether the image content determined the variation in NSS between observers, independent of the observer panel and the experimental conditions.

C.2.1 Eye-tracking Experiment

In the eye-tracking experiment in [Wang 11c] (hereafter referred to as UN), the data of NSS was collected for the twenty-nine source images of the LIVE database [Sheikh 05] by asking human observers to look freely to these images. Similar to UN, two additional eye-tracking experiments were performed: one in a laboratory at the Delft University of Technology, The Netherlands (hereafter referred to as TUD) [Liu 09], and the other one in a laboratory at the University of Western Sydney, Australia (hereafter referred to as UWS) [Engelke 09a]. The three experiments were not conducted jointly for the purpose of comparison, but rather were carried out independently using well-calibrated equipment and a well-defined experimental protocol with the aim to find "ground truth" saliency data. The details of the experiments have been introduced in Section 3. For the purpose of reminding, some important parameters for each experimental setup are introduced again here and are listed in Table C.1.

C.2.2 Variation in saliency among individuals

A fixation density map (FDM) representative for visual saliency is derived from gaze patterns recorded from eye-tracking (see [Liu 11], [Engelke 11] and [Le Meur 10b]). It is constructed by adding to each fixation location a Gaussian patch (about 2 degrees of visual angle for all three experiments), which approximates the size of the fovea in the human eye and the decrease of eye-tracker's accuracy. The intensity of the resulting saliency map ranges from 0 to 1. More details on the construction of a saliency map

	UN	TUD	UWS
Eye-tracker	SMI iView X Hi-Speed (500 Hz)	SMI iView X RED (50 Hz)	EyeTech TM3 (45 Hz)
Display	19-inch LCD (resolution: 1280×1024)	19-inch LCD (resolution: 1024×768)	19-inch LCD (resolution: 1024×768)
Participants	21 non-experts (11 males and 10 females)	20 non-experts (12 males and 8 females)	15 non-experts (9 males and 6 females)
Stimuli presentation	15s (+ 3s mid-gray screen)	10s (+ 3s mid-gray screen)	12s (+ 3s mid-gray screen)
Viewing distance	70 cm	60 cm	60 cm

Table C.1: Parameters of the setup of the eye-tracking experiments.

can be found in Section 3.2.3. In this chapter, two types of FDM are generated: one in which the mean saliency is calculated over all fixations of all subjects (MSM); and a second one, in which saliency is calculated with the fixations of an individual subject only (ISM). The variation in NSS between human observers is then quantified as the correlation coefficient between the MSM and each ISM, averaged over all participants (i.e. the ρ-value, ranging $[-1, 1]$ as used in [Liu 11]). This averaged ρ-value provides a quantitative correspondence in saliency between observers: a large value indicates a small variation in saliency among participants, while a small value indicates that the saliency is widely spread among subjects.

Figure C.1 illustrates the averaged ρ-values for the twenty-nine images of the LIVE database, computed from the three eye-tracking experiments respectively. It clearly shows that the averaged ρ-value strongly varies over the different scenes in each experiment; the difference between the highest and lowest value of ρ is 0.25 for UN, 0.30 for TUD, and 0.25 for UWS. To evaluate whether the scene dependency of ρ is equal over the three databases, and thus independent of the observer panel and of the experimental conditions, the ρ-values of the 29 images are first ranked in descending order for each laboratory independently. This results in an entity R containing three elements per image:

$$R(image) = [rank_{TUD}, rand_{UN}, rank_{UWS}]$$

where $rank_{UN}$, $rank_{TUD}$, and $rank_{UWS}$ are the actual ranks of the ρ-value per image for UN, TUD and UWS respectively. Figure C.2 illustrates the three-dimensional scatter plot of the entity R for the 29 images. This plot visualizes the consistency of the ranking of images between different experiments; the shorter the distance of a marker (i.e. an image) to the space diagonal is, the more consistent are the ranks among the three laboratories for that image. In general, all images are spread around the space diagonal

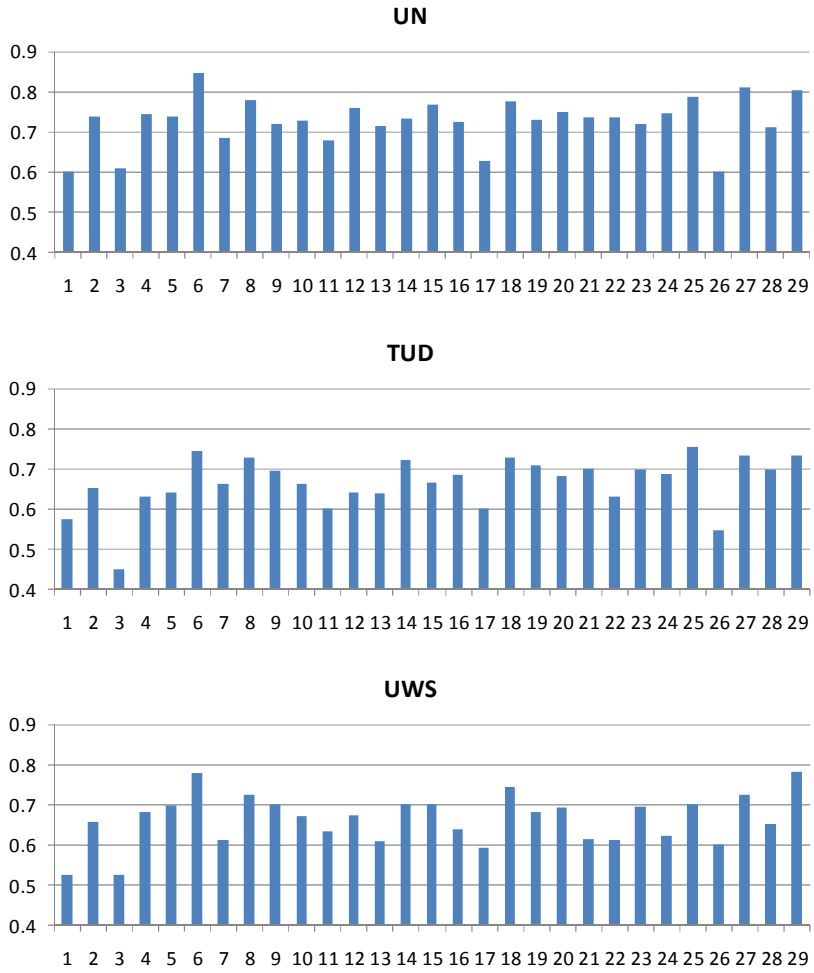


Figure C.1: Correlation coefficient (ρ-value) between the MSM and the ISM averaged overall all subjects per image, for the eye-tracking data of UN, TUD, and UWS, respectively. The vertical axis indicates the averaged ρ-value, and the horizontal axis indicates the twenty-nine images of the LIVE database.

in a rather compact manner, suggesting that ranking the images according to variation in saliency between individuals (i.e. the ρ-value) is not sensitive to the choice of subjects or experimental conditions.

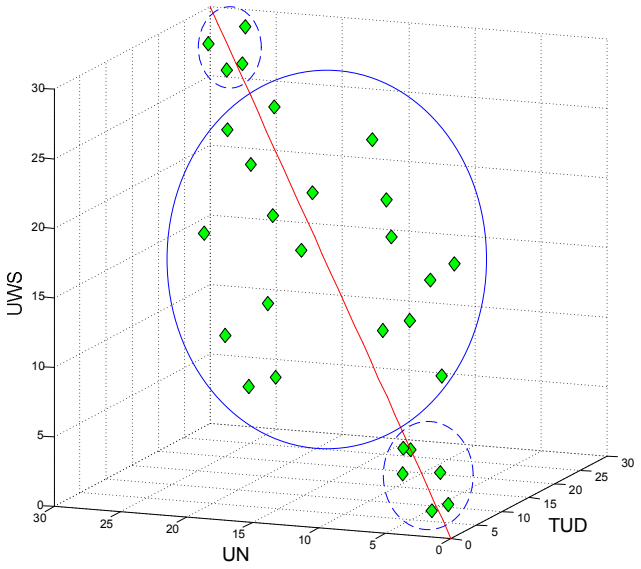


Figure C.2: The three-dimensional scatter plot of the ranks of the 29 images in the three eye-tracking experiments, i.e. UN, TUD, and UWS. Each green marker indicates one individual image, and the red line indicates the space diagonal.

C.2.3 Saliency-based content Classification

The inter-laboratory comparison demonstrates that the image content strongly, if not completely, determines the extent of variation in saliency between observers. This essentially makes the classification of image content based on the ρ-value rather meaningful. In Figure C.2, image clustering patterns are clearly observed: (1) six images are located very closely to the space diagonal in the region of high ρ-values (i.e. from rank 1 to 6 with a ρ-value higher than 0.7), (2) four images are placed closely to the space diagonal in the region of low ρ-values (i.e. from rank 26 to 29 with a ρ-value lower than 0.63), and (3) the remaining nineteen images with a medium ρ-value are spread around the space diagonal. This implies that the three eye-tracking experiments mainly share a high agreement in saliency of images with a high or low consistency in saliency between individuals.

To check the resulting classification on the LIVE database, the images with a high, medium or low ρ-value are visualized in Figure C.3. The images yielding the six highest ρ-values in each of the three eye-tracking experiments generally contain a few salient features, such as faces or texts; the saliency converges around these features for each participant. The images with the four lowest ρ-values in each of the experiments clearly lack

highly salient features, with the consequence that the saliency is randomly distributed over the whole image for the different subjects.

C.3 Impact of FDM on image quality assessment

C.3.1 Objective image quality assessment metric.

For practical reasons, the objective metrics used in our evaluation are limited to three full-reference metrics that are so far widely accepted in the image quality community: PSNR (peak signal-to-noise ratio [Wang 06]), SSIM (structural similarity index [Wang 04]) and VIF (visual information fidelity [Sheikh 06]).

PSNR

The peak signal-to-noise ratio [Wang 06] measures the difference (i.e. mean squared error), pixel by pixel, between the original and the distorted images. The mean squared error (MSE) is computed by the equation:

$$MSE = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}[I(i,j) - K(i,j)]^2$$

where I and K represent the distorted image and its original version, respectively; m and n are the resolution of the images. A low MSE value means that the two images are similar. To follow a convention that a higher value indicates greater similarity, PSNR is difined as:

$$PSNR = 10log_{10}(\frac{L^2}{MSE})$$

where L is the dynamic range of the pixel value (e.g. L=255 for grayscale images). Due to the low complexity of computation and the good performance for certain types of distortion, PSNR has been one of the most widely used metric for the purposed of assessing the quality of images and videos.

SSIM

The structural similarity index [Wang 06] computes three terms: luminance, contrast and structure. It performs the assessment of perceptual quality based on the assumption that the HVS is highly adapted for the structural information of a scene. For two images patches (which are referred to as x and y) obtained from the same location of the original and distorted images, the SSIM value is computed as:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

(a) High ρ-value images.

(b) Medium ρ-value images.

(c) Low ρ-value images.

Figure C.3: Saliency-based (i.e., on the ρ-value) image classification for the 29 source images of the LIVE database: six images had a high value of ρ in each of the three eye-tracking experiments, four images had a low value of ρ in each of the experiments, and the remaining nineteen images were classified to the category of medium ρ-value.

where μ_x, μ_y, σ_x^2, σ_y^2 and σ_{xy} represent the means, the variances and the covariance between the two image patches x and y, respectively; C_1, C_2 and C_3 ($C_3 = C_2/2$) are small constants which are included to prevent instabilities when the denominator tends to zero.

VIF

The visual information fidelity [Wang 06] is based on quantifying how much the information in the original image can be extracted from its distorted version. A scale-space-orientation wavelet is first performed using the steerable pyramid [Simoncelli 92]. Each subband in the source (i.e. the original version) is modeled as $C = S \cdot U$, where S is a random field of scalars and U is a Gaussian vector random field. The distortion model is $D = GC + \nu$, where G is a scalar gain field and ν is an additive Gaussian noise random field. An assumption is then made: the distorted and source images pass through the HVS and the uncertainty of HVS is modeled as visual noise N and N' for source and distorted image, respectively (where N and N' are zero-mean uncorrelated multivariate Gaussians). Let $E = C + N$ and $F = D + N'$, the VIF is then computed as:

$$VIF = \frac{\sum_{j \in allsubbands} I(C^j; F^j | s^j)}{\sum_{j \in allsubbands} I(C^j; E^j | s^j)}$$

where $I(X;Y|Z)$ is the conditional mutual information between X and Y, conditioned on Z; s^j is a realization of S^j for a particular image. Note that in this study, we implement the VIF in the spatial domain (i.e. pixel domain as described in [Sheikh 05]).

C.3.2 Integration of FDM into quality models

In the literature, several combination strategies of integrating saliency information into objective quality metrics have been proposed and investigated. However, comparing different combination strategies is not in the scope of the present study. In this chapter, only one combination strategy which has been proven to be efficient is used in our work. Following the procedure in [Liu 11], we integrate the saliency map into objective quality model by doing local and multiplicative weighting of the respective distortion map (as illustrated in Figure C.1). The addition of saliency to PSNR, SSIM and VIF results in three attention-based metrics, which are referred to as WPSNR, WSSIM, and WVIF. They can be defined as follows:

$$WMetric = \frac{\sum_{x=1}^{M} \sum_{y=1}^{N} [D(x,y) \cdot S_i(x,y)]}{\sum_{x=1}^{M} \sum_{y=1}^{N} S_i(x,y)}$$

where D represents the distortion map calculated by the given metric, S indicates the saliency map used, and $WMetric$ denotes the resulting attention-based metric (e.g.

WPSNR, WSSIM and WVIF).

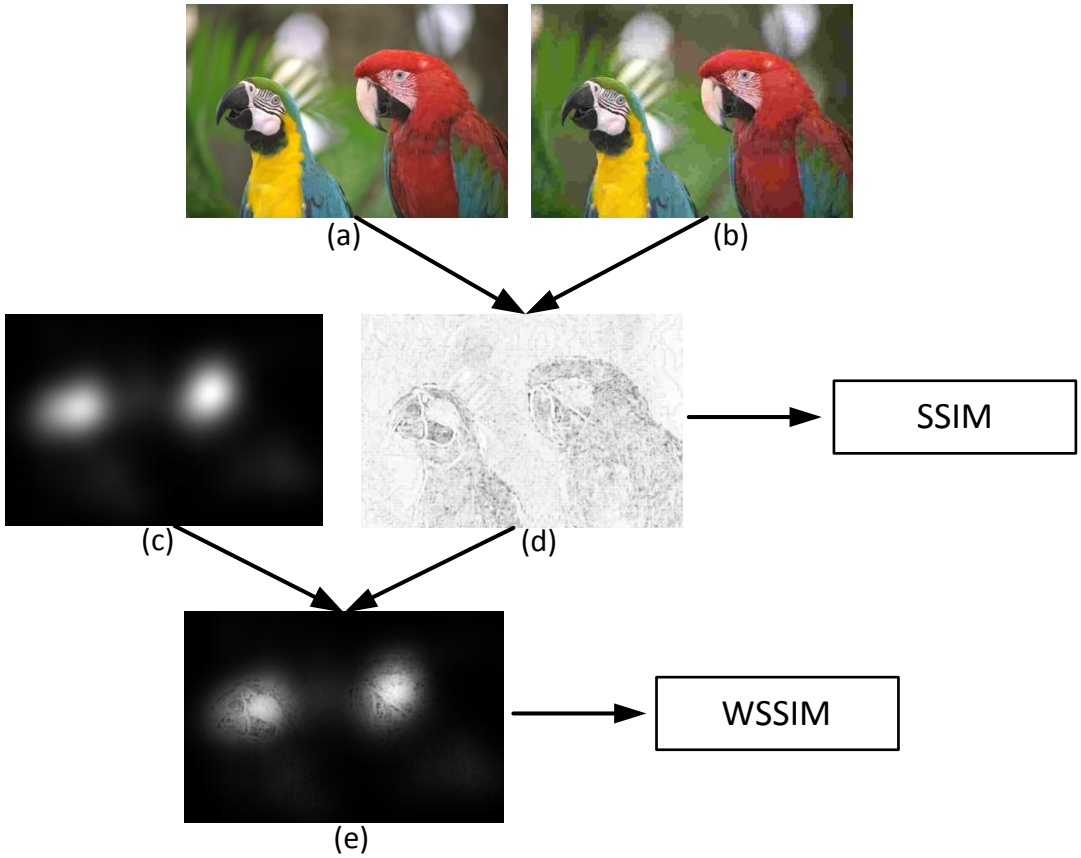


Figure C.1: Illustration of an objective metric based on saliency. (a) original image. (b) The JPEG compressed image. (c) The corresponding fixation density map showing NSS, the lower the intensity, the lower the attention. (d) Distortion map of SSIM calculated for (a) and (b). (e) Combination of distortion map and NSS, the lower the intensity, the larger the distortion is.

C.3.3 The effect of content dependency on objective image quality metrics

The added value of including visual attention in OQMs has been proven by some previous studies in the literature; including NSS obtained from eye-tracking data into several well-known OQMs in general improves the metric's prediction performance over the performance of the same metric without visual attention. This conclusion was drawn based on an evaluation over the entire LIVE database, including a variety of distortion types.

To further investigate the image content dependency in the performance gain of adding saliency to OQMs, we repeated the experiment in [Liu 11] for the source images classified as "low", "medium" and "high" ρ-value, respectively (i.e. by taking into account the extent of variation in saliency between individuals). The number of stimuli for each category and for each distortion type provided by the LIVE database is listed in Table C.2.

	JPEG	J2K	White noise	Gaussian blur	Fast fading
Total	223	227	174	174	174
Low ρ-value	34	31	24	24	24
Med. ρ-value	151	149	114	114	114
High ρ-value	48	47	36	36	36

Table C.2: Number of stimuli for each category of saliency-based image classification and for each distortion type as provided by the LIVE database.

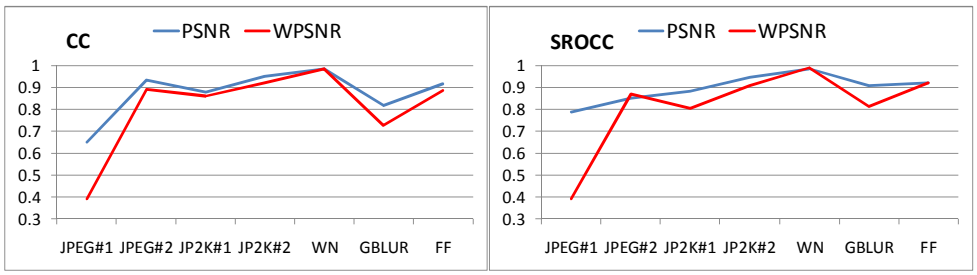
The performance of each metric is quantified by the Pearson linear correlation coefficient (CC) indicating prediction accuracy, the Spearman rank order correlation coefficient (SROCC) indicating prediction monotonicity, and the root-mean-squared error (RMSE), between the subjective scores (i.e., DMOS) and the metric's predictions.

Table C.3 and Figure C.2-C.4 illustrate the comparison in performance gain between a metric and its attention-based version averaged over all distortion types and for the images of "low", "medium" and "high" ρ-value separately. The overall gain (averaged over distortion type where appropriate) of an attention-based metric over its corresponding metric without NSS is summarized in Table C.4. Both figures and tables demonstrate that the actual gain in prediction performance that can be obtained by including saliency in objective metrics is largely affected by the ρ-value, independent of the metric used and of the image distortion type tested. In general, it shows the consistent trend that the amount of performance gain increases as the ρ-value increases. The rate of increase (i.e. from "low" to "medium" ρ-value or from "medium" to "high" ρ-value), however, depends on the metric and on the distortion type. The increase in performance gain as a consequence of the increase of ρ-value is not obvious for the subset of the LIVE database distorted by white noise (compared to other distortion types, see Figure C.2-C.4), and also is not obvious for the VIF metric (compared to other two metrics, see Table C.4). Differences in the rate of increase in performance gain may be attributed to the fact that the performance of a original metric (i.e., without NSS) varies with the distortion type, and the overall performance differs for different metrics. As such, it is more difficult to obtain a significant increase in performance by adding NSS when a high prediction performance is already achieved with a metric (e.g., a high performance is found for all three metrics applied to the "white noise" distortion, and the overall performance of VIF is already high). Such relatively small performance gains possibly mask the rate of

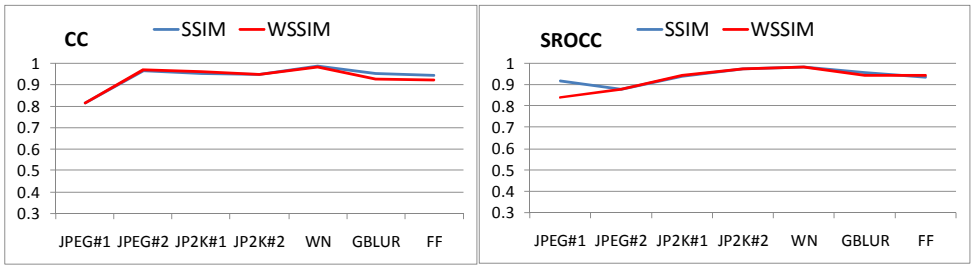
		JPEG 1	JPEG 2	J2K 1	J2K 2	White noise	Gaussian blur	Fast fading
Low ρ-value	PSNR /WPSNR	2.350 /3.145	1.504 /1.938	2.135 /2.047	0.996 /1.260	0.735 /0.733	3.460 /3.329	1.678 /1.797
	SSIM /WSSIM	1.927 /1.955	1.210 /1.215	0.972 /0.933	1.524 /1.552	0.700 /0.713	0.942 /1.071	1.092 /1.306
	VIF /WVIF	1.207 /1.179	0.788 /0.482	0.731 /1.347	1.165 /0.713	0.526 /0.468	0.896 /0.916	1.748 /1.806
Medium ρ-value	PSNR /WPSNR	1.477 /1.735	1.495 /1.419	1.318 /1.378	1.358 /1.347	0.980 /1.030	2.032 /1.847	1.456 /1.424
	SSIM /WSSIM	1.024 /0.846	1.487 /1.283	1.902 /1.721	1.648 /1.456	1.087 /0.872	1.094 /0.999	1.937 /1.739
	VIF /WVIF	1.607 /0.633	0.841 /0.578	0.794 /1.090	0.927 /0.663	0.958 /0.657	1.434 /0.680	1.641 /1.278
High ρ-value	PSNR /WPSNR	1.949 /1.621	1.532 /1.298	1.673 /1.774	1.947 /1.370	1.310 /1.315	1.998 /1.875	1.583 /1.453
	SSIM /WSSIM	1.230 /0.753	2.030 /1.532	1.185 /0.906	2.916 /2.348	1.166 /0.965	1.765 /0.818	1.526 /1.186
	VIF /WVIF	1.519 /0.636	0.595 /0.389	0.973 /0.613	1.140 /0.666	0.847 /0.778	1.030 /0.459	1.441 /1.083

Table C.3: Comparison in performance gain (quantified by RMSE) between a metric and its attention-based version (PSNR versus WPSNR, or SSIM versus WSSIM, or VIF versus WVIF) separately for the image sets of "low", "medium" and "high" ρ-value from the LIVE database.

Low ρ-value: PSNR vs. WPSNR



Low ρ-value: SSIM vs. WSSIM



Low ρ-value: VIF vs. WVIF

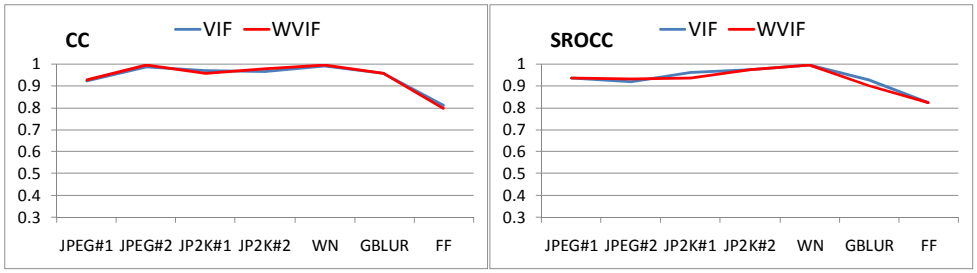
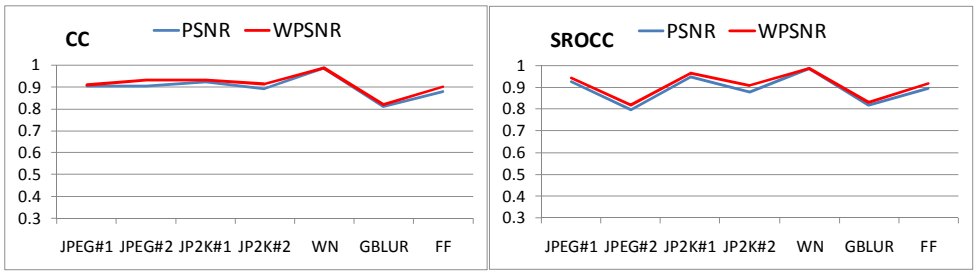
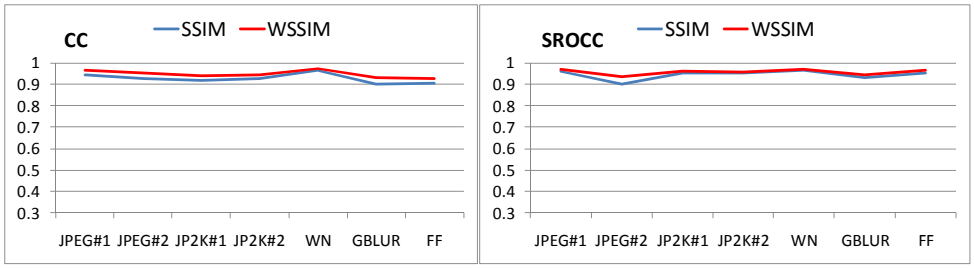


Figure C.2: Comparison in performance gain (quantified by CC and SROCC) between a metric and its attention-based version (PSNR versus WPSNR, or SSIM versus WSSIM, or VIF versus WVIF) separately for the image sets of "low" ρ-value of the LIVE database.

Mediumρ-value: PSNR vs. WPSNR



Mediumρ-value: SSIM vs. WSSIM



Mediumρ-value: VIF vs. WVIF

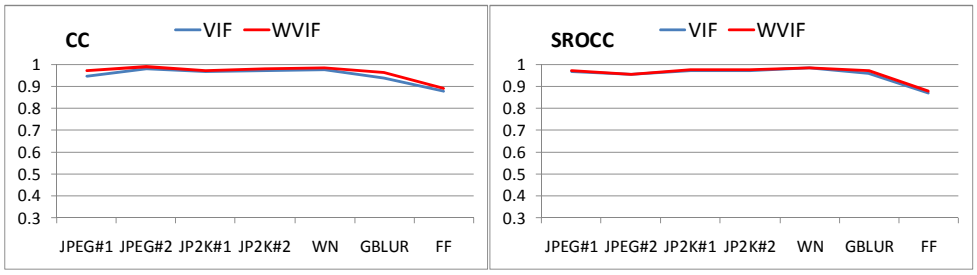
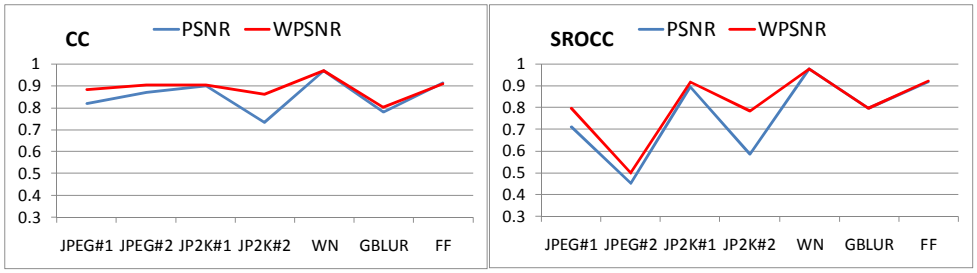
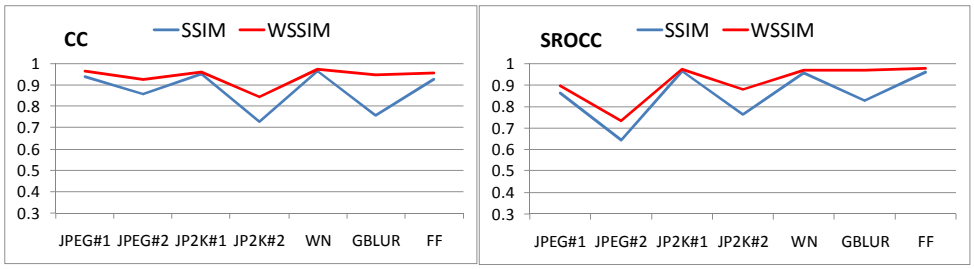


Figure C.3: Comparison in performance gain (quantified by CC and SROCC) between a metric and its attention-based version (PSNR versus WPSNR, or SSIM versus WSSIM, or VIF versus WVIF) separately for the image sets of "Medium" ρ-value of the LIVE database.

Highρ-value: PSNR vs. WPSNR



Highρ-value: SSIM vs. WSSIM



Highρ-value: VIF vs. WVIF

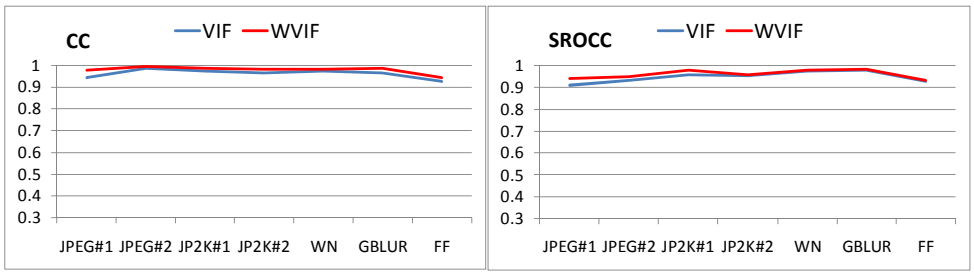


Figure C.4: Comparison in performance gain (quantified by CC and SROCC) between a metric and its attention-based version (PSNR versus WPSNR, or SSIM versus WSSIM, or VIF versus WVIF) separately for the image sets of "High" ρ-value of the LIVE database.

increase.

	|WPSNR-PSNR|			|WSSIM-SSIM|			|WVIF-VIF|	
	CC	SROCC	RMSE	CC	SROCC	RMSE	CC	SROCC	RMSE
Low ρ-value	-7%	-8%	0.20	-0.5%	-1%	0.05	0%	-0.4%	-0.02
Medium ρ-value	1%	2%	0	2%	1%	-0.18	1%	0.4%	-0.37
High ρ-value	4%	5%	-0.18	6%	6%	-0.47	2%	1%	-0.42

Table C.4: Performance gain between a metric and its attention-based version averaged over all distortion types for the image sets of "low", "medium" and "high" ρ-value from the LIVE database.

ρ-value		JPEG 1	JPEG 2	J2K 1	J2K 2	White noise	Gaussian blur	Fast fading
Low	PSNR / WPSNR	1	1	0	1	0	1	1
	SSIM / WSSIM	0	0	0	0	0	1	0
	VIF / WVIF	0	1	1	0	1	0	0
Medium	PSNR / WPSNR	1	1	1	1	1	1	1
	SSIM / WSSIM	1	1	1	1	1	1	1
	VIF / WVIF	1	1	1	1	1	1	1
High	PSNR / WPSNR	1	0	1	1	0	0	0
	SSIM / WSSIM	1	0	0	1	1	1	1
	VIF / WVIF	1	1	0	1	1	1	1

Table C.5: Results of t-test based on M-DMOS residuals: "1" means that the difference in performance between the metric with NSS and the same metric without NSS is statistically significant, and "0" means that the difference is not statistically significant.

To also check whether the numerical difference in performance between a metric with NSS and the same metric without NSS is statistically significant, hypothesis testing is conducted. The test is based on the residuals between the DMOS and the quality predicted by the metric (referred to as M-DMOS residuals). A paired-sample t-test is used to test the statistical significance for the difference between the two sets of M-DMOS residuals (i.e. one from the metric itself and one from the same metric after adding the NSS). The paired-sample t-test starts from the null hypothesis stating that the residuals of one metric are statistically indistinguishable (with 95% confidence) from the residuals of that same metric with NSS. The results of this t-test are given in Table C.5 for

all metrics and distortion types separately. This table illustrates that in most cases the difference in prediction performance by adding NSS to an objective metric is statistically significant. For the condition of "medium" ρ-value, all combinations of metrics applied to all given distortion types were tested statistically significant. It should, however, be noted that the outcome of statistical significance testing largely depends on the number of sample points, which is limited for the conditions "low" ρ-value and "high" ρ-value (see Table C.2).

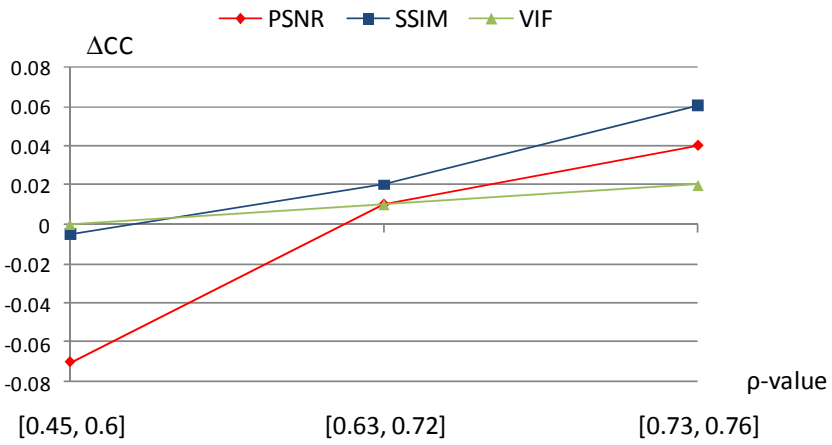


Figure C.5: Plot of the performance gain of objective metrics when adding saliency (quantified by the increase in CC, i.e. ΔCC). The horizontal axis indicates the range of ρ-value for the image sets of "low", "medium" and "high" ρ-value from the LIVE database, respectively.

Figure C.5 plots the performance gain of the OQM when saliency is added for the different ranges of ρ-value (i.e. for the "low", "medium" and "high" image sets). It demonstrates that for a ρ-value above 0.63, adding saliency is beneficial to improve the quality prediction performance of the objective metrics PSNR, SSIM and VIF. On the other hand, for the images of "low ρ-value" (i.e. $\rho \in [0.45, 0.6]$), the performance gain caused by adding saliency is either non-existing or even negative. No change in performance when adding saliency can be expected for the images, in which the saliency is uniformly spread over the entire image. To these images, including saliency to a metric does not make any difference with averaging the distortions detected over the whole image without weighting. A decrease in prediction performance by adding saliency can occur when the randomly distributed saliency coincidentally gives more weight to non-distorted regions than to heavily distorted regions in an image. For the images of "medium ρ-value" (i.e. $\rho \in [0.63, 0.72]$) and "high p-value" (i.e. $\rho \in [.0.73, 0.76]$) the performance gain is always positive. And including saliency results in a larger gain for the images of "high ρ-value" than for the images of "medium ρ-value". As illustrated in Table C.4, the gain in performance when adding saliency is twice as high for the images of "high ρ-value" than for the images of "medium ρ-value".

C.4 Conclusion

In this study, based on eye-tracking data, we found out that the extent of variation in visual attention between observers is strongly image content dependent. The smaller the variation in saliency (i.e. the more convergent the saliency map), the larger the actual performance gain that can be obtained by including saliency in objective metrics. For the images having a large variation in saliency, adding saliency to a metric runs the risk of degrading the performance of the metric for quality prediction. For the PNSR, SSIM and VIF metrics, we found a threshold value in ρ of 0.63, above which including saliency is beneficial for objective quality prediction.

Our new findings regarding to the content dependency are valuable to guide developers or users of image quality metrics to decide, based on their own application environment, whether or not to include saliency. Of course, calculating a ρ-value in order to decide whether or not to include saliency is so far unrealistic in most application context. Nevertheless, since this study illustrates that the ρ-value is linked to the spread in the saliency map, this spread may be used as an alternative decision criterion: saliency should only be applied in case this spread is low. More research needs to be done to exactly formulate the criterion in terms of spread in the (predicted) saliency map, and as such, to make the criterion practically applicable.

In addition, one should take into account the performance of a metric without saliency, since this chapter illustrates that the performance gain is limited in case the metric without saliency already has a good performance (e.g., VIF). For those metrics, computational amount can be reduced by avoiding the computation of saliency. This, in a way, also implies that the additional computational cost to include saliency may as well be used to improve the basic objective metrics' performance.

Key points

Context

- ❏ The modeling of objective quality metric (OQM) aims to automatically predict image's quality scores that are close to the scores given by human observers. A large number of OQM have been investigated in the literature, some of them focus on a particular type of distortion, while some others are designed to assess the overall perceived quality of images.
- ❏ Several of the existing OQM already have a considerably good performance. Nevertheless, researchers now attempt to further improve the prediction performance of OQM by taking into account higher level aspects of the HVS, e.g., the visual attention.
- ❏ Most of existing studies incorporate visual attention based on the assumption that a distortion occurring in a salient area is more annoying than the distortion in other areas. Nevertheless, the mechanism and the influence of visual attention for image quality assessment have not been fully understood yet.

Contributions

- ❏ We investigate the impact of adding natural scene saliency into objective quality metrics by taking into account three widely accept full-reference quality metrics: PSNR, SSIM, and VIF; the fixation density map obtained from three different eye-tracking databases are also included in our study.
- ❏ We study how the image content dependency affects the added value of natural scene saliency on image quality assessment. We found out that the extent of variation in visual attention between observers is strongly image content dependent. For the images having small variation in saliency distribution, a larger performance gain can be obtained by including saliency in objective metrics. For the images having a large variation in saliency, adding saliency to a metric runs the risk of degrading the performance of the metric for quality prediction.

Bibliography

[Achanta 09] R. Achanta, S. Hemami, F. Estrada & S. Susstrunk. *Frequency-tuned salient region detection.* In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 1597–1604. IEEE, 2009.

[Avidan 07] S. Avidan & A. Shamir. *Seam carving for content-aware image resizing.* In ACM Transactions on Graphics (TOG), volume 26, page 10. ACM, 2007.

[Barlow 67] H.B. Barlow, C. Blakemore & JD Pettigrew. *The neural mechanism of binocular depth discrimination.* The Journal of physiology, vol. 193, no. 2, page 327, 1967.

[Bichot 01] N.P. Bichot. *Attention, eye movements, and neurons: Linking physiology and behavior.* Vision and attention, pages 209–232, 2001.

[Borji 12] A. Borji & L. Itti. *State-of-the-art in Visual Attention Modeling.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012.

[Bowler 89] P.J. Bowler. Evolution: the history of an idea. Univ of California Press, 1989.

[Brookings 96] J.B. Brookings, G.F. Wilson & C.R. Swain. *Psychophysiological responses to changes in workload during simulated air traffic control.* Biological Psychology, vol. 42, no. 3, pages 361–377, 1996.

[Bruce 05] N.D.B. Bruce & J.K. Tsotsos. *An attentional framework for stereo vision.* In Computer and Robot Vision, 2005. Proceedings. The 2nd Canadian Conference on, pages 88–95. IEEE, 2005.

[Bruce 09] N.D.B. Bruce & J.K. Tsotsos. *Saliency, attention, and visual search: An information theoretic approach.* Journal of Vision, vol. 9, no. 3, 2009.

[Bruneau 02] D. Bruneau, M.A. Sasse & JD McCarthy. *The eyes never lie: The use of eye tracking data in HCI research.* In Proceedings of the CHI, volume 2, page 25, 2002.

[Byrne 99] M.D. Byrne, J.R. Anderson, S. Douglass & M. Matessa. *Eye tracking the visual search of click-down menus.* In Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit, pages 402–409. ACM, 1999.

[Cassin 90] B. Cassin, S. Solomon & M.L. Rubin. Dictionary of eye terminology. Wiley Online Library, 1990.

[Cerf 08] M. Cerf, J. Harel, W. Einhäuser & C. Koch. *Predicting human gaze using low-level saliency combined with face detection.* Advances in neural information processing systems, vol. 20, 2008.

[Cerf 09] M. Cerf, E.P. Frady & C. Koch. *Faces and text attract gaze independent of the task: Experimental data and computer model.* Journal of Vision, vol. 9, no. 12, 2009.

[Chamaret 10] C. Chamaret, S. Godeffroy, P. Lopez & O. Le Meur. *Adaptive 3D rendering based on region-of-interest.* In Proceedings of SPIE, volume 7524, page 75240V, 2010.

[Chen 10] W. Chen, J. Fournier, M. Barkowsky & P. Le Callet. *NEW REQUIREMENTS OF SUBJECTIVE VIDEO QUALITY ASSESSMENT METHODOLOGIES FOR 3DTV.* In Video Processing and Quality Metrics 2010 (VPQM), Scottsdale, USA, 2010., 2010.

[Cheng 04] H. Cheng, J.K. Barnett, A.S. Vilupuru, J.D. Marsack, S. Kasthurirangan, R.A. Applegate & A. Roorda. *A population study on changes in wave aberrations with accomodation.* Journal of Vision, vol. 4, no. 4, 2004.

[Chikkerur 10] S. Chikkerur, T. Serre, C. Tan & T. Poggio. *What and where: A Bayesian inference theory of attention.* Vision research, vol. 50, no. 22, pages 2233–2247, 2010.

[Cowen 02] L. Cowen, L.J. Ball & J. Delin. *An eye movement analysis of web page usability.* PEOPLE AND COMPUTERS, pages 317–336, 2002.

[Criminisi 03] A. Criminisi, P. Perez & K. Toyama. *Object removal by exemplar-based inpainting.* In Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, volume 2, pages II–721. IEEE, 2003.

[Davey 94] D.D. Davey, S. Naryshkin, M.L. Nielsen & T.S. Kline. *Atypical squamous cells of undetermined significance: interlaboratory comparison and quality assurance monitors.* Diagnostic cytopathology, vol. 11, no. 4, pages 390–396, 1994.

[Desimone 95] R. Desimone & J. Duncan. *Neural mechanisms of selective visual attention.* Annual review of neuroscience, vol. 18, no. 1, pages 193–222, 1995.

[Deubel 96] H. Deubel & W.X. Schneider. *Saccade target selection and object recognition: Evidence for a common attentional mechanism.* Vision research, vol. 36, no. 12, pages 1827–1837, 1996.

[Doležel 98] J. Doležel, J. Greilhuber, S. Lucretti, A. Meister, MA Lysák, L. Nardi & R. Obermayer. *Plant genome size estimation by flow cytometry: inter-laboratory comparison.* Annals of Botany, vol. 82, no. suppl 1, pages 17–26, 1998.

[Dorr 10] M. Dorr, T. Martinetz, K.R. Gegenfurtner & E. Barth. *Variability of eye movements when viewing dynamic natural scenes.* Journal of vision, vol. 10, no. 10, 2010.

[Duchowski 02] A.T. Duchowski. *A breadth-first survey of eye-tracking applications.* Behavior Research Methods, vol. 34, no. 4, pages 455–470, 2002.

[Duchowski 07] A.T. Duchowski. Eye tracking methodology: Theory and practice. Springer-Verlag New York Inc, 2007.

[Engelke 09a] U. Engelke, A. Maeder & H.J. Zepernick. *Visual attention modelling for subjective image quality databases.* In Multimedia Signal Processing, 2009. MMSP'09. IEEE International Workshop on, pages 1–6. IEEE, 2009.

[Engelke 09b] U. Engelke, H.J. Zepernick & A. Maeder. *Visual attention modeling: region-of-interest versus fixation patterns.* In Picture Coding Symposium, 2009. PCS 2009, pages 1–4. IEEE, 2009.

[Engelke 10] U. Engelke, A. Maeder & H.J. Zepernick. *Analysing inter-observer saliency variations in task-free viewing of natural images.* In Image Processing (ICIP), 2010 17th IEEE International Conference on, pages 1085–1088. IEEE, 2010.

[Engelke 11] U. Engelke, H. Kaprykowsky, H.-J. Zepernick & P. Ndjiki-Nya. *Visual Attention in Quality Assessment.* Signal Processing Magazine, IEEE, vol. 28, no. 6, pages 50 –59, nov. 2011.

[Etz 00] Stephen P. Etz & Jiebo Luo. *Ground truth for training and evaluation of automatic main subject detection.* volume 3959, pages 434–442. SPIE, 2000.

[Fawcett 06] T. Fawcett. *An introduction to ROC analysis.* Pattern recognition letters, vol. 27, no. 8, pages 861–874, 2006.

[Fincham 57] EF Fincham & J. Walton. *The reciprocal actions of accommodation and convergence.* The Journal of Physiology, vol. 137, no. 3, pages 488–508, 1957.

[Friedman 79] A. Friedman. *Framing pictures: The role of knowledge in automatized encoding and memory for gist.* Journal of Experimental Psychology: General, vol. 108, pages 316–355, 1979.

[Frintrop 06] S. Frintrop. Vocus: A visual attention system for object detection and goal-directed search, volume 3899. Springer-Verlag New York Inc, 2006.

[Gao 08] D. Gao, V. Mahadevan & N. Vasconcelos. *On the plausibility of the discriminant center-surround hypothesis for visual saliency.* Journal of vision, vol. 8, no. 7, 2008.

[Goldberg 99] J.H. Goldberg & X.P. Kotval. *Computer interface evaluation using eye movements: methods and constructs.* International Journal of Industrial Ergonomics, vol. 24, no. 6, pages 631–645, 1999.

[Goldberg 02] J.H. Goldberg, M.J. Stimson, M. Lewenstein, N. Scott & A.M. Wichansky. *Eye tracking in web search tasks: design implications.* In Proceedings of the 2002 symposium on Eye tracking research & applications, pages 51–58. ACM, 2002.

[Hakkinen 10] Jukka Hakkinen, Takashi Kawai, Jari Takatalo, Reiko Mitsuya & Gote Nyman. *What do people look at when they watch stereoscopic movies?* volume 7524, page 75240E. SPIE, 2010.

[Halle 05] M. Halle. *Autostereoscopic displays and computer graphics.* In ACM SIGGRAPH 2005 Courses, page 104. ACM, 2005.

[Hamker 05] F.H. Hamker. *The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision.* Computer Vision and Image Understanding, vol. 100, no. 1, pages 64–106, 2005.

[Hantao 11] LIU Hantao. *Modeling Perceived Quality for Imaging Applications*. PhD thesis, Delft University of Technology, The Netherlands, 2011.

[Held 10] R.T. Held, E.A. Cooper, J.F. O'Brien & M.S. Banks. *Using blur to affect perceived distance and size*. ACM Transactions on Graphics, vol. 29, 2010.

[Hoffman 08] D.M. Hoffman, A.R. Girshick, K. Akeley & M.S. Banks. *Vergence–accommodation conflicts hinder visual performance and cause visual fatigue*. Journal of Vision, vol. 8, no. 3, 2008.

[Hoffman 10] D.M. Hoffman & M.S. Banks. *Focus information is used to interpret binocular images*. Journal of vision, vol. 10, no. 5, 2010.

[Hou 07] X. Hou & L. Zhang. *Saliency detection: A spectral residual approach*. In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, pages 1–8. Ieee, 2007.

[Howard 95] I.P. Howard & B.J. Rogers. Binocular vision and stereopsis. Oxford University Press, USA, 1995.

[Huynh-Thu 11a] Q. Huynh-Thu, M. Barkowsky, P. Le Callet*et al. The Importance of Visual Attention in Improving the 3D-TV Viewing Experience: Overview and New Perspectives*. IEEE Transactions on Broadcasting, vol. 57, no. 2, pages 421–431, 2011.

[Huynh-Thu 11b] Q. Huynh-Thu & L. Schiatti. *Examination of 3D visual attention in stereoscopic video content*. In Proceedings of SPIE, volume 7865, page 78650J, 2011.

[Itti 98] L. Itti, C. Koch & E. Niebur. *A model of saliency-based visual attention for rapid scene analysis*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 20, no. 11, pages 1254–1259, 1998.

[Itti 01] L. Itti & C. Koch. *Computational modeling of visual attention*. Nature reviews neuroscience, vol. 2, no. 3, pages 194–203, 2001.

[Itti 03] L. Itti, N. Dhavale & F. Pighin. *Realistic avatar eye and head animation using a neurobiological model of visual attention*. In Proceedings of SPIE, 2003.

[Jacob 03] R.J.K. Jacob & K.S. Karn. *Eye tracking in human-computer interaction and usability research: Ready to deliver the promises*. Mind, vol. 2, no. 3, page 4, 2003.

[James 80] W. James, F. Burkhardt & I.K. Skrupskelis. The principles of psychology, volume 1. Harvard Univ Pr, 1980.

[Jansen 09] L. Jansen, S. Onat & P. König. *Influence of disparity on fixation and saccades in free viewing of natural scenes*. Journal of Vision, vol. 9, no. 1, 2009.

[Judd 09] T. Judd, K. Ehinger, F. Durand & A. Torralba. *Learning to predict where humans look*. In Computer Vision, 2009 IEEE 12th International Conference on, pages 2106–2113. IEEE, 2009.

[Judd 10] T. Judd, F. Durand & A. Torralba. *Fixations on Low Resolution Images*. Journal of Vision, vol. 10, no. 7, pages 142–142, 2010.

[Just 76] M.A. Just & P.A. Carpenter. *Eye fixations and cognitive processes*. Cognitive Psychology, vol. 8, no. 4, pages 441–480, 1976.

[Kadir 01] T. Kadir & M. Brady. *Saliency, scale and image description*. International Journal of Computer Vision, vol. 45, no. 2, pages 83–105, 2001.

[Kadiyala 08] V. Kadiyala, S. Pinneli, E. C. Larson & D. M. Chandler. *Quantifying the perceived interest of objects in images: effects of size, location, blur, and contrast*. In Thrasyvoulos N. Rogowitz Bernice E.; Pappas, editeur, Human Vision and Electronic Imaging XIII, volume 6806, pages 68060S–68060S–13, 2008.

[Karsh 83] R. Karsh & FW Breitenbach. *Looking at looking: The amorphous fixation measure*. Eye movements and psychological functions: International views, pages 53–64, 1983.

[Khan 11] R.A. Khan, E. Dinet & H. Konik. *Visual attention: Effects of blur*. In Image Processing (ICIP), 2011 18th IEEE International Conference on, pages 3289–3292. IEEE, 2011.

[Klein 00] R.M. Klein. *Inhibition of return*. Trends in cognitive sciences, vol. 4, no. 4, pages 138–147, 2000.

[Koch 85] C. Koch & S. Ullman. *Shifts in selective visual attention: towards the underlying neural circuitry*. Hum Neurobiol, vol. 4, no. 4, pages 219–27, 1985.

[Kootstra 08] G. Kootstra, A. Nederveen & B. De Boer. *Paying attention to symmetry*. In Proceedings of the British Machine Vision Conference (BMVC2008), pages 1115–1125, 2008.

[Landgrebe 06] T.C.W. Landgrebe & R.P.W. Duin. *A simplified extension of the Area under the ROC to the multiclass domain.* In 17th annual Symposium of the Pattern Recognition Association of South Africa, November 2006.

[Le Meur 06] O. Le Meur, P. Le Callet, D. Barba & D. Thoreau. *A coherent computational approach to model bottom-up visual attention.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 28, no. 5, pages 802–817, 2006.

[Le Meur 07] O. Le Meur, P. Le Callet & D. Barba. *Predicting visual fixations on video based on low-level visual features.* Vision research, vol. 47, no. 19, pages 2483–2498, 2007.

[Le Meur 09] O. Le Meur & P. Le Callet. *What we see is most likely to be what matters: visual attention and applications.* In Image Processing (ICIP), 2009 16th IEEE International Conference on, pages 3085–3088. IEEE, 2009.

[Le Meur 10a] O. Le Meur & J.C. Chevet. *Relevance of a feed-forward model of visual attention for goal-oriented and free-viewing tasks.* Image Processing, IEEE Transactions on, vol. 19, no. 11, pages 2801–2813, 2010.

[Le Meur 10b] O. Le Meur, A. Ninassi, P. Le Callet & D. Barba. *Overt visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric.* Signal Processing: Image Communication, vol. 25, no. 7, pages 547–558, 2010.

[Lee 05] K.W. Lee, H. Buxton & J. Feng. *Cue-guided search: a computational model of selective attention.* Neural Networks, IEEE Transactions on, vol. 16, no. 4, pages 910–924, 2005.

[Li 10] J. Li, Y. Tian, T. Huang & W. Gao. *Probabilistic multi-task learning for visual saliency estimation in video.* International journal of computer vision, vol. 90, no. 2, pages 150–165, 2010.

[Liu 09] H. Liu & I. Heynderickx. *Studying the added value of visual attention in objective image quality metrics based on eye movement data.* In Image Processing (ICIP), 2009 16th IEEE International Conference on, pages 3097–3100. IEEE, 2009.

[Liu 10] Y. Liu, L.K. Cormack & A.C. Bovik. *Natural scene statistics at stereo fixations.* In Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, pages 161–164. ACM, 2010.

[Liu 11] H. Liu & I. Heynderickx. *Visual Attention in Objective Image Quality Assessment: Based on Eye-Tracking Data.* Circuits and Systems for Video Technology, IEEE Transactions on, vol. 21, no. 7, pages 971–982, 2011.

[Luo 11] Y. Luo, J. Yuan, P. Xue & Q. Tian. *Saliency density maximization for object detection and localization.* Computer Vision–ACCV 2010, pages 396–408, 2011.

[Ma 02] Y.F. Ma, L. Lu, H.J. Zhang & M. Li. *A user attention model for video summarization.* In Proceedings of the tenth ACM international conference on Multimedia, pages 533–542. ACM, 2002.

[Maeder 95] A.J. Maeder. *Importance maps for adaptive information reduction in visual scenes.* In Intelligent Information Systems, 1995. ANZIIS-95. Proceedings of the Third Australian and New Zealand Conference on, pages 24–29. IEEE, 1995.

[Maki. 96a] A. Maki. *Stereo Vision in Attentive Scene Analysis. Ph. D. dissertation.* PhD thesis, Royal Inst. Tech., 1996.

[Maki 96b] A. Maki, P. Nordlund & J.O. Eklundh. *A computational model of depth-based attention.* In Pattern Recognition, 1996., Proceedings of the 13th International Conference on, volume 4, pages 734–739. IEEE, 1996.

[Maki 00] A. Maki, P. Nordlund & J.O. Eklundh. *Attentional scene segmentation: integrating depth and motion.* Computer Vision and Image Understanding, vol. 78, no. 3, pages 351–373, 2000.

[Marat 09] S. Marat, T. Ho Phuoc, L. Granjon, N. Guyader, D. Pellerin & A. Guérin-Dugué. *Modelling spatio-temporal saliency to predict gaze direction for short videos.* International Journal of Computer Vision, vol. 82, no. 3, pages 231–243, 2009.

[Marshall 96] J.A. Marshall, C.A. Burbeck, D. Ariely, J.P. Rolland & K.E. Martin. *Occlusion edge blur: a cue to relative visual depth.* JOSA A, vol. 13, no. 4, pages 681–688, 1996.

[Marshall 00] S.P. Marshall. *Method and apparatus for eye tracking and monitoring pupil dilation to evaluate cognitive activity*, July 18 2000. US Patent 6,090,051.

[Mather 96] G. Mather. *Image blur as a pictorial depth cue.* Proceedings of the Royal Society of London. Series B: Biological Sciences, vol. 263, no. 1367, pages 169–172, 1996.

[Mather 09] G. Mather. Foundations of sensation and perception, volume 2. Psychology Press, 2009.

[McAllister 93] D.F. McAllister. Stereo computer graphics and other true 3d technologies. Princeton University Press, 1993.

[Moorthy 09] A.K. Moorthy & A.C. Bovik. *Visual importance pooling for image quality assessment.* Selected Topics in Signal Processing, IEEE Journal of, vol. 3, no. 2, pages 193–201, 2009.

[Nakayama 89] K. Nakayama, S. Shimojo & G. H. Silverman. *Stereoscopic depth: its relation to image segmentation, grouping, and the recognition of occluded objects.* Perception, vol. 18, pages 55–68, 1989.

[Neri 99] P. Neri, A.J. Parker & C. Blakemore. *Probing the human stereoscopic system with reverse correlation.* Nature, vol. 401, no. 6754, pages 695–698, 1999.

[Neri 04] P. Neri, H. Bridge & D.J. Heeger. *Stereoscopic processing of absolute and relative disparity in human visual cortex.* Journal of Neurophysiology, vol. 92, no. 3, page 1880, 2004.

[Nikara 68] T. Nikara, P. O. Bishop & J. D. Pettigrew. *Analysis of retinal correspondence by studying receptive fields of rinocular single units in cat striate cortex.* Experimental Brain Research, vol. 6, pages 353–372, 1968.

[Ninassi 07] A. Ninassi, O. Le Meur, P. Le Callet & D. Barbba. *Does where you Gaze on an Image Affect your Perception of Quality? Applying Visual Attention to Image Quality Metric.* In Image Processing, 2007. ICIP 2007. IEEE International Conference on, volume 2, pages II –169 –II –172, 16 2007-oct. 19 2007.

[Okada 06] Y. Okada, K. Ukai, J.S. Wolffsohn, B. Gilmartin, A. Iijima & T. Bando. *Target spatial frequency determines the response to conflicting defocus-and convergence-driven accommodative stimuli.* Vision Research, vol. 46, no. 4, pages 475–484, 2006.

[Oliva 01] A. Oliva & A. Torralba. *Modeling the shape of the scene: A holistic representation of the spatial envelope.* International Journal of Computer Vision, vol. 42, no. 3, pages 145–175, 2001.

[Oliva 05] A. Oliva. *Gist of the scene.* In L. Itti, G. Rees & J. K. Tsotsos, editeurs, The Encyclopedia of Neurobiology of Attention, pages 251–256. Elsevier, San Diego, CA, 2005.

[Osberger 98] W. Osberger & A.J. Maeder. *Automatic identification of perceptually important regions in an image*. In Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on, volume 1, pages 701–704. IEEE, 1998.

[Ouerhani 00] N. Ouerhani & H. Hugli. *Computing visual attention from scene depth*. In Pattern Recognition, 2000. Proceedings. 15th International Conference on, volume 1, pages 375–378. IEEE, 2000.

[Palmer 08] S.E. Palmer & J.L. Brooks. *Edge-region grouping in figure-ground organization and depth perception.* Journal of Experimental Psychology: Human Perception and Performance, vol. 34, no. 6, page 1353, 2008.

[Pang 08] D. Pang, A. Kimura, T. Takeuchi, J. Yamato & K. Kashino. *A stochastic model of selective visual attention with a dynamic Bayesian network*. In Multimedia and Expo, 2008 IEEE International Conference on, pages 1073–1076, 23 2008-April 26 2008.

[Park 02] K. Park & H.W. Park. *Region-of-interest coding based on set partitioning in hierarchical trees*. Circuits and Systems for Video Technology, IEEE Transactions on, vol. 12, no. 2, pages 106–113, 2002.

[Parkhurst 02] D. Parkhurst, K. Law & E. Niebur. *Modeling the role of salience in the allocation of overt visual attention*. Vision research, vol. 42, no. 1, pages 107–123, 2002.

[Pentland 87] A.P. Pentland. *A new sense for depth of field*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, no. 4, pages 523–531, 1987.

[Perreira Da Silva 10] M. Perreira Da Silva. *Modèle computationnel d'attention pour la vision adaptative*. PhD thesis, 2010.

[Peterson 04] M.S. Peterson, A.F. Kramer & D.E. Irwin. *Covert shifts of attention precede involuntary eye movements*. Attention, Perception, & Psychophysics, vol. 66, no. 3, pages 398–405, 2004.

[Pinneli 08] S. Pinneli & D.M. Chandler. *A Bayesian approach to predicting the perceived interest of objects*. In Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on, pages 2584–2587. IEEE, 2008.

[Poggio 77] GF Poggio, B. Fischer*et al. Binocular interaction and depth sensitivity in striate and prestriate cortex of behaving rhesus monkey.* J Neurophysiol, vol. 40, no. 6, pages 1392–1405, 1977.

[Poole 06] A. Poole & L.J. Ball. *Eye tracking in HCI and usability research.* Encyclopedia of human computer interaction, pages 211–219, 2006.

[Posner 90] MI Posner & SE Petersen. *The attention system of the human brain.* Annual review of neuroscience, vol. 13, page 25, 1990.

[Potapova 11] E. Potapova, M. Zillich & M. Vincze. *Learning what matters: combining probabilistic models of 2D and 3D saliency cues.* Computer Vision Systems, pages 132–142, 2011.

[Privitera 00] C.M. Privitera & L.W. Stark. *Algorithms for defining visual regions-of-interest: Comparison with eye fixations.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 22, no. 9, pages 970–982, 2000.

[Rajashekar 08] U. Rajashekar, I. van der Linde, A.C. Bovik & L.K. Cormack. *GAFFE: A gaze-attentive fixation finding engine.* Image Processing, IEEE Transactions on, vol. 17, no. 4, pages 564–573, 2008.

[Ramasamy 09] C. Ramasamy, D.H. House, A.T. Duchowski & B. Daugherty. *Using eye tracking to analyze stereoscopic filmmaking.* In SIGGRAPH'09: Posters, page 28. ACM, 2009.

[Rayner 89] K. Rayner & A. Pollatsek. The psychology of reading. Hillsdale, NJ: Lawrence Erlbaum., 1989.

[Salvucci 98] D.D. Salvucci & J.R. Anderson. *Tracing eye movement protocols with cognitive process models.* 1998.

[Salvucci 99] D.D. Salvucci. *Mapping eye movements to cognitive processes.* PhD thesis, Carnegie Mellon University, 1999.

[Salvucci 00] D.D. Salvucci & J.H. Goldberg. *Identifying fixations and saccades in eye-tracking protocols.* In Proceedings of the 2000 symposium on Eye tracking research & applications, pages 71–78. ACM, 2000.

[Scharstein 03] D. Scharstein & R. Szeliski. *High-accuracy stereo depth maps using structured light.* In Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, volume 1, pages I–195. IEEE, 2003.

[Scharstein 07] D. Scharstein & C. Pal. *Learning conditional random fields for stereo.* In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, pages 1–8. IEEE, 2007.

[Semmlow 79] J.L. Semmlow & D. Heerema. *The role of accommodative convergence at the limits of fusional vergence.* Investigative ophthalmology & visual science, vol. 18, no. 9, pages 970–976, 1979.

[Sheikh 05] HR Sheikh, Z. Wang, L. Cormack & AC Bovik. *LIVE image quality assessment database release 2 (2005)*, 2005.

[Sheikh 06] H.R. Sheikh & A.C. Bovik. *Image information and visual quality.* Image Processing, IEEE Transactions on, vol. 15, no. 2, pages 430–444, 2006.

[Simoncelli 92] E.P. Simoncelli, W.T. Freeman, E.H. Adelson & D.J. Heeger. *Shiftable multiscale transforms.* Information Theory, IEEE Transactions on, vol. 38, no. 2, pages 587–607, 1992.

[Stankiewicz 11] Brian J. Stankiewicz, Nathan J. Anderson & Richard J. Moore. *Using performance efficiency for testing and optimization of visual attention models.* volume 7867, page 78670Y. SPIE, 2011. WWW

[Talmi 99] K. Talmi & J. Liu. *Eye and gaze tracking for visually controlled interactive stereoscopic displays.* Signal Processing: Image Communication, vol. 14, no. 10, pages 799–810, 1999.

[Tatler 07] B.W. Tatler. *The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions.* Journal of Vision, vol. 7, no. 14, 2007.

[Torralba 03] A. Torralba. *Modeling global scene factors in attention.* JOSA A, vol. 20, no. 7, pages 1407–1418, 2003.

[Treisman 80] A.M. Treisman & G. Gelade. *A feature-integration theory of attention.* Cognitive psychology, vol. 12, no. 1, pages 97–136, 1980.

[Tseng 09] P.H. Tseng, R. Carmi, I.G.M. Cameron, D.P. Munoz & L. Itti. *Quantifying center bias of observers in free viewing of dynamic natural scenes.* Journal of Vision, vol. 9, no. 7, 2009.

[Tsotsos 95] J.K. Tsotsos, S.M. Culhane, W.Y. Kei Wai, Y. Lai, N. Davis & F. Nuflo. *Modeling visual attention via selective tuning.* Artificial intelligence, vol. 78, no. 1, pages 507–545, 1995.

[Urvoy 12] M. Urvoy, M. Barkowsky, R. Cousseau, Y. Koudota, V. Ricorde, P. Le Callet, J. Gutiérrez & N. García. *NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences.* In Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on, pages 109–114. IEEE, 2012.

[Vijayakumar 01] S. Vijayakumar, J. Conradt, T. Shibata & S. Schaal. *Overt visual attention for a humanoid robot.* In Intelligent Robots and Systems, 2001. Proceedings. 2001 IEEE/RSJ International Conference on, volume 4, pages 2332–2337. IEEE, 2001.

[Vu 03] K. Vu, K.A. Hua & W. Tavanapong. *Image retrieval based on regions of interest.* Knowledge and Data Engineering, IEEE Transactions on, vol. 15, no. 4, pages 1045–1049, 2003.

[Walther 06] D. Walther & C. Koch. *Modeling attention to salient proto-objects.* Neural Networks, vol. 19, no. 9, pages 1395–1407, 2006.

[Wandell 95] B.A. Wandell. Foundations of vision, volume 21. Sinauer Associates, 1995.

[Wang 04] Z. Wang, A.C. Bovik, H.R. Sheikh & E.P. Simoncelli. *Image quality assessment: From error visibility to structural similarity.* Image Processing, IEEE Transactions on, vol. 13, no. 4, pages 600–612, 2004.

[Wang 06] Z. Wang & A.C. Bovik. *Modern image quality assessment.* Synthesis Lectures on Image, Video, and Multimedia Processing, vol. 2, no. 1, pages 1–156, 2006.

[Wang 10] Junle Wang, Damon M. Chandler & Patrick Le Callet. *Quantifying the relationship between visual salience and visual importance.* In Bernice E. Rogowitz & Thrasyvoulos N. Pappas, editeurs, Human Vision and Electronic Imaging XV - part of the IS&T-SPIE Electronic Imaging Symposium, San Jose, CA, USA, January 18-21, 2010, Proceedings, volume 7527 of *SPIE Proceedings*, page 75270. SPIE, 2010.

[Wang 11a] D. Wang, G. Li, W. Jia & X. Luo. *Saliency-driven scaling optimization for image retargeting.* The Visual Computer, pages 1–8, 2011.

[Wang 11b] J. Wang, P Le Callet, V. Ricordel & S Tourancheau. *Quantifying depth bias in free viewing of still stereoscopic synthetic stimuli.*

16th European Conference on Eye Movements, Marseille, France, 2011.

[Wang 11c] Pepion R. Wang J. & P. Le Callet. *IRCCyN/IVC eyetracker images LIVE database*, 2011.

[Watt 05] S.J. Watt, K. Akeley, M.O. Ernst & M.S. Banks. *Focus cues affect perceived depth.* Journal of Vision, vol. 5, no. 10, 2005.

[Wedel 07] M. Wedel & R. Pieters. *A review of eye-tracking research in marketing.* Review of marketing research, vol. 4, pages 123–147, 2007.

[Werlberger 09] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers & H. Bischof. *Anisotropic Huber-L1 optical flow.* In Proceedings of the British machine vision conference, 2009.

[Werlberger 10] M. Werlberger, T. Pock & H. Bischof. *Motion estimation with non-local total variation regularization.* In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pages 2464–2471. IEEE, 2010.

[Wheatstone 38] C. Wheatstone. *Contributions to the physiology of vision.–Part the first. On some remarkable, and hitherto unobserved, phenomena of binocular vision.* Philosophical transactions of the Royal Society of London, vol. 128, pages 371–394, 1838.

[Widdel 84] H Widdel. *Operational problems in analysing eye movements.* Advances in Psychology, pages 21–29, 1984.

[Wismeijer 10] DA Wismeijer, CJ Erkelens, R. van Ee & M. Wexler. *Depth cue combination in spontaneous eye movements.* Journal of vision, vol. 10, no. 6, 2010.

[Wolfe 00] J.M. Wolfe. *Visual attention.* Seeing, vol. 2, pages 335–386, 2000.

[Wolfe 04] J.M. Wolfe & T.S. Horowitz. *What attributes guide the deployment of visual attention and how do they do it?* Nature Reviews Neuroscience, vol. 5, no. 6, pages 495–501, 2004.

[Yarbus 67] A.L. Yarbus. Eye movements and vision. Plenum press, 1967.

[Zhang 08] L. Zhang, M.H. Tong, T.K. Marks, H. Shan & G.W. Cottrell. *SUN: A Bayesian framework for saliency using natural statistics.* Journal of Vision, vol. 8, no. 7, 2008.

[Zhang 10] Y. Zhang, G. Jiang, M. Yu & K. Chen. *Stereoscopic visual attention model for 3D video.* Advances in Multimedia Modeling, pages 314–324, 2010.

[Zhao 11] Q. Zhao & C. Koch. *Learning a saliency map using fixated locations in natural scenes.* Journal of Vision, vol. 11, no. 3, 2011.

l'université nantes angers le mans
PÔLE DE RECHERCHE ET D'ENSEIGNEMENT SUPÉRIEUR

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Junle Wang

From 2D to Stereoscopic-3D Visual Saliency: Revisiting Psychophysical Methods and Computational Modeling

Saillance Visuelle, de la 2D à la 3D Stéréoscopique : Examen des Méthodes Psychophysique et Modélisation Computationnelle

Résumé

L'attention visuelle est l'un des mécanismes les plus importants mis en oeuvre par le système visuel humain (SVH) afin de réduire la quantité d'information que le cerveau a besoin de traiter pour appréhender le contenu d'une scène. Un nombre croissant de travaux est consacré à l'étude de l'attention visuelle, et en particulier à sa modélisation computationnelle. Dans cette thèse, nous présentons des études portant sur plusieurs aspects de cette recherche. Nos travaux peuvent être classés globalement en deux parties. La première concerne les questions liées à la vérité de terrain utilisée, la seconde est relative à la modélisation de l'attention visuelle dans des conditions de visualisation 3D.

Dans la première partie, nous analysons la fiabilité de cartes de densité de fixation issues de différentes bases de données occulométriques. Ensuite, nous identifions quantitativement les similitudes et les différences entre carte de densité de fixation et carte d'importance visuelle, ces deux types de carte étant les vérités de terrain communément utilisées par les applications relatives à l'attention. Puis, pour faire face au manque de vérité de terrain exploitable pour la modélisation de l'attention visuelle 3D, nous procédons à une expérimentation oculométrique binoculaire qui aboutit à la création d'une nouvelle base de données avec des images stéréoscopiques 3D.

Dans la seconde partie, nous commençons par examiner l'impact de la profondeur sur l'attention visuelle dans des conditions de visualisation 3D. Nous quantifions d'abord le " biais de profondeur " lié à la visualisation de contenus synthétiques 3D sur écran plat stéréoscopique. Ensuite, nous étendons notre étude avec l'usage d'images 3D au contenu naturel. Nous proposons un modèle de l'attention visuelle 3D basé saillance de profondeur, modèle qui repose sur le contraste de profondeur de la scène. Deux façons différentes d'exploiter l'information de profondeur par notre modèle sont comparées. Ensuite, nous étudions le biais central et les différences qui existent selon que les conditions de visualisation soient 2D ou 3D. Nous intégrons aussi le biais central à notre modèle de l'attention visuelle 3D. Enfin, considérant que l'attention visuelle combinée à une technique de floutage peut améliorer la qualité d'expérience de la TV-3D, nous étudions l'influence de flou sur la perception de la profondeur, et la relation du flou avec la disparité binoculaire.

Mots clés

Attention visuelle, modèle computationnel, occulométrie, 3DTV, saillance visuelle, importance visuelle, expérience subjective, stéréoscopie.

Abstract

Visual attention is one of the most important mechanisms deployed in the human visual system to reduce the amount of information that our brain needs to process. An increasing amount of efforts are being dedicated in the studies of visual attention, particularly in computational modeling of visual attention. In this thesis, we present studies focusing on several aspects of the research of visual attention. Our works can be mainly classified into two parts. The first part concerns ground truths used in the studies related to visual attention ; the second part contains studies related to the modeling of visual attention for Stereoscopic 3D (S-3D) viewing condition.

In the first part, our work starts with identifying the reliability of FDM from different eye-tracking databases. Then we quantitatively identify the similarities and difference between fixation density maps and visual importance map, which have been two widely used ground truth for attention-related applications. Next, to solve the problem of lacking ground truth in the community of 3D visual attention modeling, we conduct a binocular eye-tracking experiment to create a new eye-tracking database for S-3D images.

In the second part, we start with examining the impact of depth on visual attention in S-3D viewing condition. We firstly introduce a so-called “depth-bias” in the viewing of synthetic S-3D content on planar stereoscopic display. Then, we extend our study from synthetic stimuli to natural content S-3D images. We propose a depth-saliency-based model of 3D visual attention, which relies on depth contrast of the scene. Two different ways of applying depth information in S-3D visual attention model are also compared in our study. Next, we study the difference of center-bias between 2D and S-3D viewing conditions, and further integrate the center-bias with S-3D visual attention modeling. At the end, based on the assumption that visual attention can be used for improving Quality of Experience of 3D-TV when collaborating with blur, we study the influence of blur on depth perception and blur's relationship with binocular disparity.

Key Words

Visual attention, computational model, eye-tracking, 3DTV, visual saliency, visual importance, subjective experiment, stereoscopy.